TOWARDS THE DIFFERENTIAL EVALUATION OF
PARKINSON'S DISEASE BY MEANS OF VOICE AND
SPEECH PROCESSING

LAUREANO MORO VELÁZQUEZ, MSc.

Advisor:

JUAN IGNACIO GODINO LLORENTE, PhD.

Thesis submitted to opt for the degree of *Doctor of Philosophy* in:
Systems and Services Engineering for the Information Society
Ingeniería de Sistemas y Servicios para la Sociedad de la Información

Escuela Técnica Superior de Ingeniería y Sistemas de
Telecomunicación

Universidad Politécnica de Madrid

May 2018

Laureano Moro Velázquez, MEng.

*Towards the differential evaluation of Parkinson's Disease by means of voice
and speech processing*

PRESENTED TO OPT FOR THE PHD. DEGREE IN:

Systems and Services Engineering for the Information Society

Ingeniería de Sistemas y Servicios para la Sociedad de la Información

INSTITUTION:

E. T. S. I. S. T.

Escuela Técnica Superior de Ingeniería y Sistemas de Telecomunicación

Universidad Politécnica de Madrid

ADVISOR:

Juan Ignacio Godino Llorente, PhD.

Madrid

May 2018

TITLE OF THE THESIS:

Towards the differential evaluation of Parkinson's Disease by means of voice and speech processing

AUTHOR:

Laureano Moro Velázquez, MSc.

ADVISOR:

Juan Ignacio Godino Llorente, PhD.

COMMITTEE MEMBERS:

Francisco Grandas Pérez, PhD.	<i>Hospital General Universitario Gregorio Marañón</i>
Javier Macías Guarasa, PhD.	<i>Universidad de Alcalá</i>
Jiří Mekyska, PhD.	<i>Brno University of Technology</i>
Rubén San Segundo Hernández PhD.	<i>Universidad Politécnica de Madrid</i>
José Luis Blanco Murillo, PhD.	<i>Universidad Politécnica de Madrid</i>

EXTERNAL REVIEWERS:

Virginia Yagüe Jiménez, PhD.	<i>Consejo Superior de Investigaciones Científicas</i>
Jesús Antonio Villalba , PhD.	<i>Johns Hopkins University</i>
Julián David Arias Londoño, PhD.	<i>Universidad de Antioquia</i>

Madrid, on the 8th of June, 2018. After the defense of the PhD dissertation at Escuela Técnica Superior de Ingeniería y Sistemas de Telecomunicación of the Universidad Politécnica de Madrid, the board agrees to grant the following grade:

The chair

The secretary

The members

Dedicado a mis padres, Tere y Antonio.

*"Si el niño no estuviera tan profundamente dormido
sentiría en su moflete de nardo la lágrima resbalada
desde la vieja mejilla de cuero."*

— José Luis Sampedro, *La sonrisa etrusca*

ABSTRACT

Parkinson's Disease (PD) is a neurodegenerative condition that affects to 1% of population over the age of 60 in industrialized countries. This disease seriously affects a patient's independence and motor capabilities, having a considerable impact on their daily activities as it advances. Early detection can potentially slow the progression of PD but, unfortunately, the required period of clinical diagnosis ranges from months to years. Therefore, new objective and reliable tools are needed to support the differential diagnosis of the disease and to reduce this time.

The analysis of a patient's speech has demonstrated to provide relevant information about the presence of the disease and, consequently, is a possible source of features to be used in diagnosis systems. Speech, an ability that is almost universal, involves coordination and precision of movements in mainly the laryngeal and articulatory muscles. The purpose of this thesis is to propose and study different approaches to support clinical diagnosis of PD employing speech as the object of analysis.

In this thesis, five sets of experiments are carried out, each one containing new approaches aimed to detect the presence of the disease in the speech of idiopathic PD patients and controls from three different databases. Four of these experiments focus on the analysis of articulatory aspects while the fifth employs phonatory features and a final combination of phonatory and articulatory information into a single approach. In these approaches, several state-of-the-art speaker and speech recognition technologies are employed in a different scenario: the automatic detection of PD from speech.

Several known feature families such as MFCC, PLP or LPC and new features based on the use of the modulation spectra are analyzed. Moreover, different speech frame selection techniques are proposed, such as allophonic distillation and acoustic landmark distillation, providing certain specific speech segments that are of interest to the purposes of this work. The main classification techniques employed are GMM-UBM and *i-Vectors*-GPLDA, along with new schemes such as the forced GMM.

As a consequence of the analysis of the proposed approaches, the influence of PD in these specific segments is examined, allowing to extract conclusions about the functioning of the parkinsonian dysarthric speech. These segments are phonetic groups related to the narrowing of the vocal tract or the use of the glottal source, such as fricatives, plosives or vowels, or include mainly transitions between phonetic units, such as the beginning of a burst or the end of a vowel, more related with the coordination of the articulators.

The best accuracy results in the detection of PD achieved with the proposed methodologies reach values ranging from 85% to 94% with Area Under the Curve between 0.91 and 0.99 depending on the database. These results are obtained largely by employing approaches based on the frame selection proposed techniques: allophonic and acoustic landmark distillations.

In the same manner, it is concluded that the discriminatory properties of the proposed phonatory approaches to automatically detect PD are quite limited in comparison with the analyzed articulatory approaches. Results suggest that PD affects the movements related to all of the studied articulatory segmental groups but has a clearer influence in the consonants with a greater narrowing of the vocal tract, mainly plosives and fricatives.

Finally, the new proposed methodologies demonstrate their ability to support the differential diagnosis of PD during a patient's clinical assessment and are a step forward in the speech-based diagnosis systems for PD employing articulatory aspects.

RESUMEN

La Enfermedad de Párkinson (EP) es un trastorno degenerativo que afecta a un 1% de la población con más de 60 años en países industrializados. Esta enfermedad afecta a la independencia del paciente y a sus capacidades motoras, teniendo un impacto considerable en sus actividades diarias a medida que dicho trastorno avanza. Una detección precoz puede potencialmente frenar la progresión de la EP pero el tiempo requerido para realizar un diagnóstico clínico puede ir de unos meses a varios años. Así pues, se hacen necesarias nuevas herramientas objetivas y consistentes que ayuden al diagnóstico diferencial de la enfermedad y reduzcan dicho tiempo.

La literatura ha demostrado que el análisis del habla de los pacientes proporciona información relevante sobre la presencia de la enfermedad y es una posible fuente de indicadores para sistemas de diagnóstico. El habla, una habilidad casi universal, conlleva la realización de movimientos precisos y coordinados, en los músculos laríngeos y articulatorios. Por lo tanto, el propósito de esta tesis es la propuesta y análisis de diferentes esquemas de ayuda al diagnóstico de la EP mediante el uso del habla como objeto de análisis.

En esta tesis, cinco sets de experimentos han sido llevados a cabo, cada uno de ellos incluyendo nuevas aproximaciones al problema destinadas a evaluar la presencia de la enfermedad en pacientes con párkinson idiopático y sujetos de control de tres bases de datos diferentes. Cuatro de estos sets analizan aspectos articulatorios mientras que el quinto emplea características fonatorias y una combinación de estas junto con las articulatorias en un único esquema. En estas aproximaciones, varias técnicas de reconocimiento de locutor y del habla se emplean en un escenario distinto: la detección de la EP a partir del habla.

Varias familias de características conocidas como los MFCC, PLP o LPC junto con otras nuevas basadas en el espectro de modulación son analizadas. Además, distintas técnicas de selección de fragmentos del habla son propuestas, como la destilación alofónica o de hitos acústicos, que permiten la obtención de ciertos segmentos del habla que son de interés para los propósitos de este trabajo. Las principales técnicas de clasificación empleadas son GMM-UBM e *i-Vectors*-GPLDA junto con otras nuevas como los GMM forzados.

A través del análisis de los esquemas propuestos, se examina la influencia de la EP en diferentes segmentos del habla, permitiendo la extracción de conclusiones sobre el funcionamiento del habla disártrica parkinsoniana. Estos segmentos están referidos a grupos fonéticos relacionados con el estrechamiento del tracto vocal o el uso de la fuente glótica o bien incluyen transiciones entre unidades fonéticas como el

inicio de una oclusión o el final de una vocal, más relacionados con la coordinación de los órganos articuladores.

Los mejores resultados de precisión en la detección de la EP obtenidos con las metodologías propuestas alcanzan valores de entre el 85% y el 94%, con *Area Under the Curve* entre 0.91 y 0.99, dependiendo de la base de datos de estudio. Estos resultados se obtienen empleando el esquema basado en las técnicas de selección de fragmentos propuestas: destilación alofónica y de hitos acústicos.

Del mismo modo, se concluye que las propiedades discriminativas del esquema fonatorio propuesto para la detección automática de la EP son limitadas en comparación con los esquemas articulatorios analizados. Los resultados sugieren que la EP afecta a los movimientos relacionados con todos los grupos de segmentos articulatorios pero tiene una influencia más clara en la actividad motora asociada a las consonantes con mayor estrechamiento del tracto vocal, principalmente, oclusivas y fricativas.

Finalmente, las nuevas metodologías propuestas pueden contribuir al diagnóstico diferencial de EP durante la evaluación clínica de pacientes y son un paso adelante para los sistemas de diagnóstico de la EP basados en los aspectos articulatorios del habla.

PUBLICATIONS

Journal papers

[i] Moro-Velazquez, L., Gomez-Garcia, J. A., Godino-Llorente, J. I., Villalba, J., Orozco-Arroyave, J. R., and Dehak, N., Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease. *Applied Soft Computing*, 62, 649-666. 2018.

[ii] Godino-Llorente J. I., Shattuck-Hufnagel S., Choi, J. Y., Moro-Velazquez, L., Gomez-Garcia J.A., Towards the identification of Idiopathic Parkinson's Disease from the speech. New articulatory kinetic biomarkers. *PLoS ONE* 12(12): e0189583, 2017.

[iii] Moro-Velazquez, L., Gomez-Garcia, J. A., Godino-Llorente, J. I., Voice Pathology Detection Using Modulation Spectrum-Optimized Metrics. *Frontiers in bioengineering and biotechnology*, 4, 2016.

[iv] Gomez-García, J. A., Moro-Velazquez, L., Godino-Llorente, J. I., Castellanos-Dominguez, C. G., An insight to the automatic categorization of speakers according to sex and its application to the detection of voice pathologies: A comparative study. *Revista Facultad de Ingeniería Universidad de Antioquia*, (79), 50-62, 2016.

[v] Moro-Velazquez, L., Gomez-Garcia, J. A., Godino-Llorente, J. I., Andrade-Miranda, G., Modulation spectra morphological parameters: a new method to assess voice pathologies according to the GRBAS scale. *BioMed research international*, 2015.

[vi] Andrade-Miranda, G., Godino-Llorente, J. I., Moro-Velazquez, L., Gomez-Garcia, J. A., An automatic method to detect and track the glottal gap from high speed videoendoscopic images. *Biomedical engineering online*, 14(1), 100, 2015.

Conference papers

[vii] Moro-Velazquez, L. Gomez-Garcia, J. A., Godino-Llorente, J. I., Rusz, J., Skodda, S., Grandas-Perez, F., Velazquez, J. M., Orozco-Arroyave, J. R., Noth, E., and Dehak, N., Study of the automatic detection of Parkinson's Disease based on speaker recognition technologies and allophonic distillation. *Engineering in Medicine and Biology Society (EMBC), 2018 40th Annual International Conference of the IEEE.*, 2018.

[viii] Moro-Velazquez, L., Godino-Llorente, J. I., Gómez-García, J. A., Villalba, J., Shattuck-Hufnagel, S., and Dehak, N., (2017). Use of acoustic landmarks and GMM-UBM blend in the automatic detection of Parkinson's disease. *10th international workshop on models and analysis of vocal emissions for biomedical applications (MAVEBA)*, 2017.

[ix] Gomez-Garcia, J.A., Moro-Velazquez, L., Mendes-Laureano, J., Godino-Llorente, J. I., (2017). On the design of a voice pathology assessment system based on the GRB scale. *10th international workshop on models and analysis of vocal emissions for biomedical applications (MAVEBA)*, 2017.

[x] Gomez-Garcia, J. A., Moro-Velazquez, L., Orozco-Arroyave, J.R., Godino-Llorente, J.I. (2016). Detection of Parkinson's disease by means of GMM-UBM and iVectors techniques. *XXXI Simposium Nacional de la Union Científica Internacional de Radio, URSI*, 2016.

[xi] Moro-Velazquez, L., Gomez-Garcia, J. A., and Godino-Llorente, J. I., Tuning of modulation spectrum parameters for voice pathology detection. *9th international workshop on models and analysis of vocal emissions for biomedical applications (MAVEBA)*, 2015.

[xii] Gomez-Garcia, J. A., Moro-Velazquez, L., Godino-Llorente, J. I. and Castellanos-Domínguez, G., Automatic age detection in normal and pathological voice. *Interspeech*, 2015.

[xiii] Moro-Velazquez, L., Gomez Garcia, J.A. and Godino-Llorente, J. I., Analysis of Complexity and Modulation Spectra parameterizations to characterize Voice Roughness. *VIII Jornadas en Tecnologías del habla, Iberspeech*, 2014.

[xiv] Moro-Velazquez, L. and Godino-Llorente, J. I., Analisis de metodos de parametrizacion para la simulacion de un sistema de evaluacion perceptual de voces patologicas. *I Jornadas Multidisciplinares de Usuarios de la Voz, el Habla y el Canto*, 2013.

[xv] Moro-Velazquez, L. and Godino-Llorente J. I., Eficiencia y acuerdo en varios metodos para la simulacion de evaluacion perceptual en voces patologicas. *VII Jornadas de Reconocimiento Biométrico de Personas*, 2013.

AGRADECIMIENTOS

Doy las gracias a mi director de tesis, el Dr. Juan Ignacio Godino, por su generosidad a la hora de transmitir conocimientos, tanto científicos como de otros ámbitos académicos igualmente importantes. Admiro tu paciencia, tu constancia y tu infatigable labor investigadora, la cual es una inspiración para mí. Gracias por tu cercanía, tu entrega y por haberme apoyado tanto en mi salto (inesperado) a Baltimore.

Gracias a mi compañero de laboratorio Jorge Andrés Gómez por tantas horas compartidas en las que hacíamos evolucionar nuestro trabajo juntos a través de reuniones espontáneas que eran casi terapéuticas. Trabajar contigo ha sido y será muy importante para mí. Conozco poca gente con tanta entrega como la tuya y deseo que la vida te premie con el éxito que mereces. Gracias por tu apoyo dentro y fuera del ámbito científico y por escucharme cuando me ha hecho falta, que últimamente ha sido en muchas ocasiones. Cuida de Paquito mientras yo no esté y recuerda: no le digas a nadie que guardamos cosas importantes en su calavera¹.

Thanks to Dr. Najim Dehak for accepting me in his laboratory, giving me the opportunity to work in the Center for Speech and Language Processing, one of the most important speech processing groups of the world. Thank you, Najim, for believing in me. Thanks to all the colleagues in the laboratory for their help and support and to all of the American family that I have found in Baltimore, specially to Tyler, without whom I would have never made it.

Muchas gracias a los doctores Virginia Yagüe, Jesús Villalba y Julián David Arias por sus comentarios a la hora de revisar la tesis y por su desinteresado y valioso trabajo.

Así mismo quiero agradecer a todos los pacientes del Hospital Gregorio Marañón y a los sujetos de control que desinteresadamente han contribuido con sus voces a la creación del corpus de voz y habla parkinsoniana Neurovoz. También agradezco la labor de los servicios de Neurología y Otorrinolaringología de dicho hospital por su colaboración en dicha tarea.

Thanks to Dr. Jan Rusz for sharing his unvaluable database of newly diagnosed patients and matched controls.

Ahora no solo quiero dar las gracias a las personas que más dentro llevo dentro de mí y a las que más presentes tengo en todo momento: mi familia y mis amigos. También les quiero pedir perdón por no haber estado ahí el último año. No os imagináis cuánto os he extrañado y os sigo extrañando. No sé cuánto durará mi viaje pero espero volver pronto.

---

¹ Vaya, ya no podemos guardarlas ahí.

Amigos, los de GETA, los erasmus, los del pueblo, los de Soge, los del roko... estoy deseando pasar todo el tiempo que pueda con vosotros y seguir creciendo a vuestro lado. Os lo he dicho unas cuantas veces, pero aquí va una más: estoy enamorado de vosotros. Me apetece un cocido.

Las familias permanecen unidas cuando todos quieren que así sea y yo tengo mucha suerte de que en la mía siempre hayamos estado juntos y apretados. Siempre recordaré que de niño me sentaba a mirar a través de la ventana del comedor, esperando con toda la ilusión del mundo a que llegasen a casa tío Jose o la yaya, que ya no están, junto con todos los demás, para celebrar un cumpleaños o lo que fuese. Esos eran los momentos más importantes de mi niñez. Niños de la familia, Alonso, Clara, Inés, Juan Manuel y Sara, haré todo lo posible por daros lo que vuestros padres y abuelos me han dado a mí. Gracias familia, sois lo más importante.

Y para finalizar, doy las gracias a las personas más importantes de mi vida: mis padres, Tere y Antonio y mi hermana Ali. Gracias por quererme, apoyarme en todo y por vuestra ayuda, la que veo y la que no. Os quiero.

CONTENTS

I INTRODUCTION

1	INTRODUCTION	3
1.1	Motivation	3
1.2	Introduction to the document	5
2	PARKINSON'S DISEASE	7
2.1	Description and pathophysiology	7
2.2	Etiology	7
2.3	Symptomatology	8
2.4	Epidemiology and prevalence	10
2.5	Diagnosis and rating scales	10
2.6	Treatment	14
2.7	Parkinson's Disease and speech	15
3	HYPOTHESIS AND GOALS	19
3.1	Hypothesis	19
3.2	Goals and objectives	19
4	STATE OF THE ART	21
4.1	Introduction	21
4.2	Phonatory aspects	24
4.2.1	Evidences of the influence of PD in the phonatory system	24
4.2.2	Acoustic measurements and perceptual evaluations	25
4.2.3	Formant measurements and associated features	27
4.2.4	Automatic detection and assessment	28
4.3	Articulatory aspects	29
4.3.1	Evidences of the influence of PD in phoneme articulation	29
4.3.2	Kinematic considerations	31
4.3.3	Formant measurements and associated features	32
4.3.4	Automatic detection and assessment	33
4.4	Prosodic aspects	34
4.4.1	Evidences of the influence of PD in prosody. The main cues: fundamental frequency and speech SPL	34
4.4.2	Speech rate in parkinsonian speakers	37
4.4.3	Other prosodic indicators	38
4.5	Linguistic aspects	39
4.5.1	Disfluency	40
4.5.2	Use of action-verbs	40
4.5.3	Other approaches	41
4.6	Combined approaches	42
4.7	Analysis of databases found in literature	44
4.8	List of features found in literature	47
4.9	Discussion	47

4.9.1	Phonatory approaches	49
4.9.2	Articulatory approaches	49
4.9.3	Prosodic approaches	50
4.9.4	Linguistic approaches	51
4.9.5	Combined approaches	51
4.9.6	Final considerations	52
5	OVERALL METHODOLOGY AND MATERIALS	55
5.1	Articulatory approaches	55
5.1.1	Experimental set 1: use of state-of-the-art speaker recognition technologies	56
5.1.2	Experimental set 2: forced Gaussian models	56
5.1.3	Experimental set 3: allophonic distillation	56
5.1.4	Experimental set 4: acoustic landmarks	57
5.2	Phonatory approaches	58
5.3	Materials: Speech databases	58
5.3.1	GITA database	58
5.3.2	Neurovoz database	61
5.3.3	CzechPD database	64
5.3.4	Albayzin database	65
5.3.5	Vystadial-Czech database	66
5.3.6	FisherSP database	66
5.4	General considerations	66
5.5	Discussion	68
II	ARTICULATORY APPROACHES	
6	EXPERIMENTAL SET 1: SPEAKER RECOGNITION TECHNOLOGIES	73
6.1	Introduction	73
6.2	Theoretical Background	75
6.2.1	Parameterizations	75
6.2.2	Kinetic changes	78
6.2.3	Classification schemes: GMM-UBM and <i>i-Vectors</i>	80
6.2.4	ROCCH-DET curves	83
6.3	Experimental Setup	83
6.3.1	Acoustic material	84
6.3.2	Methodology	85
6.4	Results	90
6.4.1	Influence of τ_{frame} , n_0 and F in the first stage of methodology	90
6.4.2	GMM-UBM and <i>i-Vectors</i> global results	93
6.4.3	Validation and cross-database trials	95
6.5	Discussion	96
6.5.1	Influence of the family of features.	98
6.5.2	Influence of τ_{frame} and $\tau_{derivative}$	100
6.5.3	Influence of kinetic changes.	102
6.5.4	Influence of speech tasks.	103
6.5.5	Influence of classification techniques.	103

6.5.6	UPDRS and H-Y assessment	104
6.5.7	Other considerations.	104
6.6	Summary	107
7	EXPERIMENTAL SET 2: FORCED GAUSSIANS	109
7.1	Introduction	109
7.2	Theoretical background	109
7.2.1	Phonetic considerations	109
7.2.2	Speech forced alignment	110
7.2.3	Forced Gaussian Mixture Model	113
7.2.4	Gaussian Mixture Model - forced Universal Back- ground Model (GMM-fUBM)	115
7.2.5	Forced Gaussian Mixture Model - forced Univer- sal Background Model (fGMM-fUBM)	115
7.2.6	Optimized forced Gaussian Mixture Model - forced Universal Background Model (fGMM-fUBM)	117
7.3	Methodology	118
7.3.1	Forced Alignment	118
7.3.2	Forced Gaussian Mixture models	119
7.4	Results	120
7.5	Discussion	121
7.6	Summary	125
8	EXPERIMENTAL SET 3: ALLOPHONIC DISTILLATION	127
8.1	Introduction	127
8.2	Theoretical background	128
8.2.1	Allophonic distillation	128
8.3	Methodology	131
8.4	Results	135
8.5	Discussion	139
8.5.1	Allophonic distillations	139
8.5.2	Analysis of approaches	143
8.5.3	Speech tasks	144
8.5.4	Other considerations	145
8.6	Summary	145
9	EXPERIMENTAL SET 4: ACOUSTIC LANDMARKS	147
9.1	Introduction	147
9.2	Theoretical background	148
9.2.1	Acoustic landmarks	148
9.2.2	Acoustic landmark distillation	150
9.3	Experimental setup	150
9.4	Results	153
9.5	Discussion	155
9.5.1	Types of landmarks	159
9.5.2	Analysis of approaches and speech tasks	162
9.5.3	Other considerations	163
9.6	Summary	163

III PHONATORY APPROACHES

10 EXPERIMENTAL SET 5: PHONATORY FEATURES	167
10.1 Theoretical background	170
10.1.1 Jitter, shimmer, noise and complexity features	170
10.1.2 Modulation Spectrum	172
10.2 Methodology	177
10.2.1 First approach: noise, jitter, shimmer and complexity	177
10.2.2 Second approach: Modulation Spectrum features	178
10.2.3 Third approach: Rasta-PLP	180
10.2.4 Fourth approach: all features	180
10.3 Results	182
10.4 Discussion	183
10.5 Summary	184

IV DISCUSSION, FUTURE WORK AND FINAL CONCLUSIONS

11 DISCUSSION	187
11.1 Experimental sets' summary	187
11.2 Best results overview	188
11.3 Discussion	189
11.3.1 The glass ceiling in the automatic detection of PD from speech	191
11.3.2 Features and kinetic changes	193
11.3.3 Analysis of experimental sets	194
11.3.4 Speech tasks	196
11.3.5 Scores and rating scales	197
11.3.6 Other considerations	197
12 CONTRIBUTIONS, FUTURE WORK AND CONCLUSIONS	199
12.1 Contributions	199
12.2 Future work	200
12.3 Conclusions	202

V APPENDIX

A APPENDIX A	207
B APPENDIX B	211

BIBLIOGRAPHY	215

LIST OF FIGURES

Figure 2.1	The balance of causal and protective environmental factors weighted by the causal and protective genetic factors in the PD occurrence. Extracted from [12]	8
Figure 5.1	Scheme of the overview of methodology employed in the thesis, including the four experiments on the articulatory approaches (a), the phonatory approaches, and the combination of the phonatory approach with one articulatory approach (b) where the articulatory part corresponds to the methodology providing the best results among the four articulatory experiments.	59
Figure 5.2	UPDRS, H-Y and years since diagnosis histograms for the PD patients within the GITA database . .	61
Figure 5.3	UPDRS, H-Y and years since diagnosis histograms for the PD patients within the Neurovoz database	62
Figure 5.4	Image employed in monologue task in Neurovoz. Speakers were asked to describe the image during, at least, 30 s.	63
Figure 5.5	UPDRS, H-Y and Years since diagnosis histograms for the PD patients within the CzechPD database	65
Figure 5.6	General procedure followed in the thesis to analyze the different approaches.	67
Figure 6.1	Waveforms and spectrograms of a newly diagnosed patient and a control speaker during a DDK task (three repetitions of /pa-ta-ka/). Obtained from Neurovoz database.	74
Figure 6.2	Waveforms and spectrograms of an advanced-stage patient and a control speaker during a DDK task (three repetitions of /pa-ta-ka/). Obtained from Neurovoz database.	75
Figure 6.3	Overview of methodology.	76
Figure 6.4	LPC, MFCC and PLP calculation procedure scheme.	78
Figure 6.5	Scheme illustrating $\tau_{derivative}$ and its relationship with τ_{frame} and the number of coefficients used to calculate Δ for $n_0 = 2$	80
Figure 6.6	First derivative of the fourth PLP coefficient ($F = 12$) calculated using two different values of n_0 during two /pa-ta-ka/ utterances of a control (upper) and a parkinsonian patient (down). . .	80
Figure 6.7	Example of DET and ROCCH-DET curves for two different sets of data.	83

Figure 6.8	Pseudocode of Methodology. First stage. *Sustained vowel was used only in this stage. ** n_0 does not always range between 1 and 7; in some cases these values were restricted to limit $\tau_{derivative}$ between 35 and 80 ms	86
Figure 6.10	Pseudocode of Methodology. Second stage. . . .	87
Figure 6.9	GMM-UBM training methodology outline. Features can be MFCC, Rasta-PLP and LPC and derivative coefficients $\Delta + \Delta\Delta$	87
Figure 6.11	<i>i-Vectors</i> training methodology outline. Features can be MFCC, Rasta-PLP and LPC with derivative coefficients $\Delta + \Delta\Delta$	88
Figure 6.12	Best accuracies in the range $\{10, \dots, 40\}$ ms of τ_{frame} for the three different parameterizations, GMM-UBM classification techniques and using speech tasks: a) DDK, b) Sentence 1, c) Sentence 2, d) Sustained vowel. These values are obtained using different number of G , F and $\tau_{derivative}$ at each point.	90
Figure 6.13	Best accuracies using $\tau_{frame} = 10$ ms as a function of $\tau_{derivative}$ for the three different parameterizations using speech tasks: a) DDK, b) Sentence 1, c) Sentence 2. These values are obtained using different number of G and F at each point.	91
Figure 6.14	Best accuracies using $\tau_{frame} = \{10, 15, 20\}$ ms as a function of $\tau_{derivative}$ for DDK, Sentence 1 and Sentence 2 using Rasta-PLP+ $\Delta + \Delta\Delta$ coefficients. These values are obtained using different number of G and F at each point.	92
Figure 6.15	Best accuracies using $\tau_{frame} = 10$ ms as a function of the number of coefficients for the three different parameterizations, GMM-UBM classification techniques and using speech tasks: a) DDK, b) Sentence 1, c) Sentence 2, d) Sustained vowel. These values are obtained using different number of G and $\tau_{derivative}$ at each point.	92
Figure 6.16	DET curves for MFCC, Rasta-PLP and LPC parameterizations + $\Delta + \Delta\Delta$ leading to the highest accuracy using the two classification techniques in Sentence 2 speech task tests.	93
Figure 6.17	Best accuracies in GMM and <i>i-Vectors</i> tests using $\tau_{frame} = 15$, $n_0 = 2$ with Rasta-PLP+ $\Delta + \Delta\Delta$ for Sentence 2 as a function of G and F . In the case of <i>i-Vectors</i> , $h = 50$ and $q = 14$. Best points of operation are marked with a red dot.	94

Figure 6.18	DET curves for the different speech tasks leading to the highest accuracies and the configurations detailed in Table 6.4.	94
Figure 6.19	Mutual information of Rasta-PLP (R-PLP) features with respect to the control/pathological classes	96
Figure 6.20	DET curves for the three parkinsonian databases and the two classification schemes. Albayzin is used as UBM for Neurovoz and GITA trials and Vystadial-Czech for CzechPD trials.	99
Figure 6.21	ROCCH-DET curves for the three parkinsonian databases and the two classification schemes. Albayzin is used as UBM for Neurovoz and GITA trials and Vystadial-Czech for CzechPD trials.	100
Figure 6.22	DET curves for the three parkinsonian databases and the two classification schemes. Albayzin is used as UBM in all the cases. Only DDK task is employed	101
Figure 6.23	ROCCH-DET curves for the three parkinsonian databases and the two classification schemes. Albayzin is used as UBM in all the cases. Only DDK task is employed	101
Figure 6.24	Best accuracies for PLP parameterization with and without Rasta filtering, using DDK speech task and GMM-UBM classification technique. τ_{frame} is varied in the range $\{10, \dots, 30\}$ ms	102
Figure 7.1	Waveforms and spectrograms of a parkinsonian (newly diagnosed) and a control speaker pronouncing the syllable /pe/. Obtained from Neurovoz database.	110
Figure 7.2	Waveforms and spectrograms of a parkinsonian (in an advanced stage) and a control speaker pronouncing the word "es". Obtained from Neurovoz database.	111
Figure 7.3	Example of forced alignment labeling obtained with Kaldi, including waveform, spectrogram, phonemes and transcription	112
Figure 7.4	fGMM general scheme. Initially, the fGMM model is trained using the training data (training fold) and then, tested with the testing data (testing fold)	115
Figure 7.5	GMM-fUBM general scheme. Initially, the model is trained using the training data (training fold) and then, tested with the testing data (testing fold)	116
Figure 7.6	fGMM-fUBM general scheme. Initially, the model is trained using the training data (training fold) and then, tested with the testing data (testing fold).	117

Figure 7.7	ROCCH-DET curves obtained employing testing scores from GITA in the classification schemes: fGMM, GMM-fUBM, fGMM-fUBM and optimized fGMM-fUBM. Fixed text is used as speech task.	121
Figure 7.8	ROCCH-DET curves obtained employing testing scores from Neurovoz in the classification schemes: fGMM, GMM-fUBM, fGMM-fUBM and optimized fGMM-fUBM. Fixed text is used as speech task.	122
Figure 7.9	Per frame sensitivity and specificity in GITA database using fixed text in an optimized fGMM-fUBM scheme.	123
Figure 7.10	Per frame sensitivity and specificity in Neurovoz database using fixed text in an optimized fGMM-fUBM scheme.	124
Figure 7.11	ROCCH-DET curves obtained employing testing scores GITA, Neurovoz and CzechPD in the GMM-fUBM classification scheme. DDK sequence (/pa-ta-ka/) is used as speech task.	124
Figure 8.1	Waveforms and spectrograms of a parkinsonian (newly diagnosed) and a control speaker pronouncing the word "petaca". Obtained from the Neurovoz database.	128
Figure 8.2	Waveforms and spectrograms of a parkinsonian (in an advanced stage) and a control speaker pronouncing the word "petaca". Obtained from Neurovoz database.	129
Figure 8.3	Example of forced alignment labeling obtained with Kaldi, including waveform, spectrogram, phonemes, phoneme category and transcription. LIQ stands for liquid; VOW, for vowel; PLOS, for plosive, FRIC, for fricative; and NAS, for nasal.	131
Figure 8.4	First approach followed for the allophonic distillation methodology.	132
Figure 8.5	Second approach followed for the allophonic distillation methodology.	133
Figure 8.6	Third approach followed for the allophonic distillation methodology.	134
Figure 8.7	ROCCH-DET curves for GITA and Neurovoz including the scores of the best results of each approach for these databases. In all the cases, fixed text is acoustic material, as these produce the best results.	136

Figure 8.8	ROCCH-DET curves for the three parkinsonian speech databases in the third approach (allophonic distillation applied only on UBM database, Albayzin). Fixed text are used for GITA and Neurovoz database, and DDK for CzechPD.	139
Figure 8.9	ROCCH-DET curves for the three parkinsonian speech databases used for testing in the cross-database trials, employing the third approach (allophonic distillation only in UBM, Albayzin.) In all cases, DDK tasks from the parkinsonian databases are used.	141
Figure 8.10	Best accuracies for the three different approaches (marked as 1, 2 or 3) and speech tasks, where ft stands for fixed text, ddk, Diadochokinetic task and monol., monologues.	142
Figure 8.11	Best AUC for the three different approaches (marked as 1, 2 or 3) and speech tasks, where ft stands for fixed text, ddk, Diadochokinetic task and monol., monologues.	143
Figure 9.1	Spectrogram and associated landmarks in the pronunciation of the sentence "how do we define it?". Red lines represent b-lmk, light and dark blue, g-lmk (pointing the beginning and end of vocal fold vibration, respectively) and green lines, s-lmk. Additionally, black lines mark the vowel landmark, in the middle of the two g-lmk and grey lines, the presence of a glide. The use of the signs +/− indicates that the transition is referred to the beginning (+) or the end (−) of an allophone. Obtained from [234].	148
Figure 9.2	Landmark preliminary candidate detection scheme based on the abruptness calculation.	150
Figure 9.3	Segments selected after acoustic landmark distillation process. When a candidate landmark is detected, only the features from the frame containing it and the two adjacent are employed in the PD detector.	151
Figure 9.4	First approach for PD detection based on the detection of landmarks.	152
Figure 9.5	Second approach for PD detection based on the detection of landmarks.	152
Figure 9.6	Third approach for PD detection based on the detection of landmarks.	153
Figure 9.7	Probability density functions of landmarks' PoC for fixed texts in GITA.	155
Figure 9.8	Probability density functions of landmarks' PoC for fixed texts in Neurovoz.	155

Figure 9.9	ROCCH-DET curves for the three parkinsonian speech databases in the third approach (acoustic landmark detection used only on UBM database, Albayzin). Fixed text are used for GITA and Neurovoz databases, and DDK for CzechPD.	159
Figure 9.10	ROCCH-DET curves for the three parkinsonian speech databases used as testing set in a cross-database scheme employing the third approach (acoustic landmark detection used only on UBM database, Albayzin). Only DDK tasks are utilized.	160
Figure 9.11	Best accuracies for the three different approaches (marked as 1, 2 or 3) and speech tasks, where ft stands for fixed text, ddk, Diadochokinetic task and monol., monologues.	161
Figure 9.12	Best AUC for the three different approaches (marked as 1, 2 or 3) and speech tasks, where ft stands for fixed text, ddk, Diadochokinetic task and monol., monologues.	161
Figure 10.1	Waveform and spectrogram of four patients during the sustained phonation of vowel /a:/. Obtained from Neurovoz database.	168
Figure 10.2	Waveform and spectrogram of four patients during the sustained phonation of vowel /a:/. Obtained from Neurovoz database.	169
Figure 10.3	Jitter and shimmer phenomena reflected in a sustained vowel signal.	171
Figure 10.4	MS of a two patient and two control speakers during the sustained phonation of vowel /a:/. Obtained from Neurovoz database.	173
Figure 10.5	Points of the MS matrix used to obtain MSH (i), MSW (ii) and LMR (iii).	175
Figure 10.6	Diagram of first approach in phonatory experimental set.	178
Figure 10.7	Diagram of second approach in the phonatory experimental set.	179
Figure 10.8	Diagram of third approach in the phonatory experimental set.	180
Figure 10.9	Diagram of the fourth approach in the phonatory experimental set.	181
Figure 10.10	ROCCH-DET curves of the combination of articulatory and phonatory approaches.	183
Figure 11.1	ROCCH-DET curves for the three parkinsonian speech databases employing the best set-ups for each experiment set. Fixed texts are selected for GITA and Neurovoz whereas DDK task from CzechPD is used.	189

Figure 11.2	ROCCH-DET curves for the three parkinsonian speech databases used for testing in the cross-database trials, employing the best set-ups for each experiment set. In all cases, DDK tasks from the parkinsonian databases are used.	192

LIST OF TABLES

Table 2.1	Grouping of Clinical Features According to Diagnostic Utility. Extracted from [19]	12
Table 2.2	Proposed Diagnostic Criteria for PD. Extracted from [19].	13
Table 4.1	Parkinsonian speech databases identified in literature.	45
Table 4.2	Some of the speech motor feature families found in literature with its behavior and works including them.	48
Table 5.1	Transcription of the six GITA fixed sentences (Spanish) and its translation to English	60
Table 5.2	Demographic statistics of GITA database. Ages are expressed in years.	60
Table 5.3	Average duration (s) of the speech tasks in GITA	60
Table 5.4	Spanish transcription of the six Neurovoz fixed sentences and its translation to English	63
Table 5.5	Demographic statistics of Neurovoz database. Ages are expressed in years.	64
Table 5.6	Average duration (s) of the speech tasks in Neurovoz.	64
Table 5.7	Demographic statistics of CzechPD database. Ages are expressed in years.	65
Table 5.8	Average duration (s) of the speech tasks in CzechPD. Ages are expressed in years.	66
Table 5.9	Summary of databases employed in this thesis. .	68
Table 6.1	Speech tasks and phonetic transcription (IPA) from GITA database. Sust. v. stands for sustained vowel	84
Table 6.2	Number of variations of each degree of freedom (Speech task, τ_{frame} , n_0 , F , G , q and h) in the two classification schemes and the three feature families	89
Table 6.3	Best accuracy results $\pm$ CI for the different acoustic materials using MFCC, Rasta-PLP and LPC with derivative coefficients $\Delta + \Delta\Delta$	93

Table 6.4	Best results for each acoustic material.	95
Table 6.5	Results using Rasta-PLP and the configuration leading to best results with and without derivatives to analyze its influence in accuracy, AUC, specificity and sensitivity.	95
Table 6.6	Results using GITA, Neurovoz and CzechPD in a GMM-UBM classification scheme, employing Albayzin for the UBM.	97
Table 6.7	Results using CzechPD within a GMM-UBM classification scheme, employing Vystadial-Czech for the UBM.	97
Table 6.8	Results using GITA, Neurovoz and CzechPD in a <i>i-Vectors</i> -GPLDA classification scheme, employing Albayzin for the UBM.	97
Table 6.9	Results using CzechPD within a <i>i-Vectors</i> -GPLDA classification scheme, employing Vystadial-Czech for the UBM.	97
Table 6.10	Cross-database results using GITA, Neurovoz and CzechPD in a GMM-UBM classification scheme, employing Albayzin for the UBM. For every trial, two parkinsonian databases are used for training and the remaining, for testing.	98
Table 6.11	Cross-database results using GITA, Neurovoz and CzechPD in a <i>i-Vectors</i> -GPLDA classification scheme, employing Albayzin for the UBM. For every trial, two parkinsonian databases are used for training and the remaining, for testing.	98
Table 6.12	Results using PLP feature family with and without Rasta filter in order to its influence in accuracy, AUC, specificity and sensitivity.	101
Table 7.1	fGMM results	120
Table 7.2	GMM-fUBM results	120
Table 7.3	fGMM-fUBM results	120
Table 7.4	Optimized fGMM-fUBM results	121
Table 7.5	Cross-database GMM-fUBM results using DDK task. For each test database, the remaining two parkinsonian databases are used to train the model.	123
Table 8.1	Categorizations of allophones suggested in [222]	130
Table 8.2	Total number of repetitions of consonants in GITA and Neurovoz fixed texts.	132
Table 8.3	Approach 1: results after applying allophonic distillation to the parkinsonian database and no distillation to the Albayzin database (UBM). Only fixed text is used as speech task.	135
Table 8.4	Approach 2: results after applying allophonic distillation to the parkinsonian and Albayzin databases (UBM). Only fixed text is used as speech task.	136

Table 8.5	Approach 3: results after applying allophonic distillation to the Albayzin database (UBM). GITA (with no allophonic distillation) is used to train and test the models.	137
Table 8.6	Approach 3: results after applying allophonic distillation to the Albayzin database (UBM). Neurovoz (with no allophonic distillation) is used to train and test the models.	137
Table 8.7	Approach 3: results after applying allophonic distillation to the Albayzin database (UBM). CzechPD (with no allophonic distillation) is used to train and test the models. Only DDK task is considered.	138
Table 8.8	Best results after the fusion of scores for the three parkinsonian databases separately. The scores were obtained after applying the third approach (allophonic distillation only in the UBM database)	138
Table 8.9	Cross-database results in GITA, Neurovoz and CzechPD, employing Albayzin for the UBM with the five different types of allophonic distillation. For every trial, two parkinsonian databases are used for training and the remaining, for testing.	140
Table 9.1	Different frequency bands in landmark detection	149
Table 9.2	First approach results for the three types of landmarks and databases using Albayzin for the UBM.	154
Table 9.3	Second approach results for the three types of landmarks and databases using Albayzin for the UBM.	156
Table 9.4	Third approach results for the three types of landmarks and databases using Albayzin for the UBM.	157
Table 9.5	Fusion of scores in the third approach for the three parkinsonian databases using Albayzin as UBM. For the sake of simplicity, the table includes only the combination of scores (2- and 3- <i>tuples</i>) leading to the best results.	158
Table 9.6	Cross-database results using DDK task and the third approach (in which the landmarks are only used in the UBM database. For every trial, two parkinsonian databases are used for training and the remaining, for testing.	158
Table 10.1	Optimum configuration to calculate MS features, as detailed in [243]. All values are in Hz.	179
Table 10.2	Best results in the four approaches for GITA database.	182
Table 10.3	Best results in the four approaches for Neurovoz database.	182

Table 10.4	Results of combination of phonatory and articulatory approaches in GITA and Neurovoz. (*) the p-value associated to these correlation values was equal or below 0.05	182
Table 11.1	List of experimental sets. Exp. set stands for experimental set and Distill. stands for distillation.	188
Table 11.2	Best results for each experiment as a function of the employed database and speech task. Sust. v. stands for sustained vowel, and Exp. for Experimental set. * indicates p-Values equal or under 0.05.	190
Table 11.3	Best cross-database results for all the experiments. In all cases, speech task is DDK.	191
Table A.1	Places of articulation, and number of joint vowels and consonants in Sentence 1	208
Table A.2	Places of articulation, and number of joint vowels and consonants in Sentence 2	209
Table B.1	Scores per speaker from the testing utterances in the approaches leading to best results of each experiment. Only fixed text from GITA database is considered.	212
Table B.2	Scores per speaker from the testing utterances in the approaches leading to best results of each experiment. Only fixed text from Neurovoz database is considered.	213
Table B.3	Scores per speaker from the testing utterances in the approaches leading to best results of each experiment. Only Diadochokinetik (DDK) task from Czech database is considered.	214

ACRONYMS

AUC	Area Under the Curve
ApEn	Approximate Entropy
BBE	Band Bark Energies
CAPE-V	Consensus Auditory Perceptual Evaluation of Voice
CI	Confidence Interval
CIL	Cumulative Intersection Level
CSD	Chi-Squared Distance
D_2	Correlation Dimension

DBS	Deep Brain Stimulation
DCT	Discrete Cosine Transform
DDK	Diadochokinetik
DET	Detection Error Trade-off
DFA	Detrented Fluctuation Analysis
DNN	Deep Neural Network
EER	Equal Error Rate
F ₂	second formant
FAM	Forced Alignment Model
FCR	Formant Centralization Ratio
FDA-2	Frenchay Dysarthria Assessment
FDLP	Frequency Domain Linear Prediction
FFT	Fast Fourier Transform
F _{tr}	Fundamental frequency tremor
FIR	Finite Impulse Response
fGMM	Forced Gaussian Mixture Model
fGMM-fUBM	Forced Gaussian Mixture Model - forced Universal Background Model
FIR	Finite Impulse Response
fMLLR	Feature space Maximum Likelihood Linear Regression
fUBM	Forced Universal Background Model
FuzzyEn	Fuzzy Entropy
GKE	Gaussian Kernel Entropy
GMM	Gaussian Mixture Model
GMM-fUBM	Gaussian Mixture Model - forced Universal Background Model
GMM-UBM	Gaussian Mixture Model-Universal Background Model
GNE	Glottal to Noise Excitation ratio
GPLDA	Gaussian Probability Linear Discriminant Analysis
GRBAS	Grade, Roughness, Breathiness, Asthenia and Strain scale
H-Y	Hoehn & Yahr
HD	Huntington's Disease
HE	Hurst Exponent

HMM	Hidden Markov Models
HNR	Harmonic to Noise Ratio
IFCC	Instantaneous Frequencies Cepstral Coefficients
IPA	International Phonetic Alphabet
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LLE	Largest Lyapunov Exponent
LMR	Low Modulation Ratio
LOO	Leave One Out
LPC	Linear Predictive Coding
LRKK ₂	Leucine-Rich Repeat Kinase 2
LSVT	Lee Silvermant Voice Treatment
MAP	Maximum a Posteriori
MDVP	Multidimensional Voice Program
MFCC	Mel-Frequency Cepstral Coefficients
MIM	Mutual Information Maximization
MLLT	Maximum Likelihood Linear Transform
MPT	Maximum Phonation Time
MRI	Magnetic Resonance Imaging
mRMR	minimum Redundancy Maximum Relevance
MS	Modulation Spectrum
mSaEn	modified Sample Entropy
MSH	Modulation Spectrum Homogeneity
MSP	Modulation Spectrum Percentiles
MSW	Modulation Spectrum Contrast
NHR	Noise to Harmonics Ratio
NLP	Natural Language Processing
NNE	Normalized Noise Energy
PCA	Principal Components Analysis
PD	Parkinson's Disease
PE	Permutation Entropy
PLP	Perceptual Linear Prediction

PoC	Probability of Candidate
PPE	Pitch Period Entropy
RALA	Ratio of Points Above Linear Average
RFF	Relative Fundamental Frequency
ROC	Receiver Operating Characteristic
ROCCH	Convex Hull around Receiver Operating Characteristic
RPDE	Recurrence Period Density Entropy
SaEn	Sample Entropy
SNR	Signal to Noise Ratio
SPI	Soft Phonation Index
SPL	Sound Pressure Level
sTFT	short-Time Fourier Transform
SVM	Support Vector Machine
SVR	Support Vector Regression
TBI	Traumatic Brain Injury
UBM	Universal Background Model
UPDRS	Unified Parkinson's Disease Rating Scale
VAI	Vowel Articulation Index
VOT	Voice Onset Time
tVSA	triangular Vowel Space Area
VSA	Vowel Space Area

Part I

INTRODUCTION

INTRODUCTION

Filling a glass of water, communicating by phone or starting to walk when the red light turns green in a crosswalk are simple and feasible actions for the vast majority of people. However, these tasks involve the coordination of a large number of muscles and can become almost impossible to achieve for some people living with a neurodegenerative disease. Ordinary tasks that certain persons suffering from Parkinson's Disease (PD) cannot perform without the help of a care giver, becoming highly dependent. That is PD, a condition whose symptoms are more complex, diverse and unknown for most of the people than a limb tremor. An accurate definitive diagnosis can take years since there are not precise biomarkers or tests to detect PD, like a brain Magnetic Resonance Imaging (MRI) or a blood test. On the other hand, although there is currently no cure for this disease, it is expected that new drugs will be able to stop its advance or, at least, to slow it down enough to extend a patient's period of independence significantly. And when this moment arrives, early diagnosis will be crucial. This thesis is aimed to contribute with the creation of new tests to support the diagnosis of PD and to reduce the time required to reach a reliable verdict.

Since PD is a motor system disorder, the analysis of a patient's movements while performing a complex motor task can lead to the identification of potential biomarkers. But, which type of tasks can be used within the analysis? Speech, being an ability almost universal, involves coordination and precision of movements in mainly the laryngeal and articulatory muscles. Therefore, the purpose of this thesis is to study different approaches to support the clinical diagnosis and evaluation of PD by employing speech as the object of observation.

1.1 MOTIVATION

PD affects to 1% of the population over the age of 60 in industrialized countries and with increasing life expectancy, it is expected to affect to more than 9 million people in industrialized nations by 2030 [1]. A slower advance of the disease in patients will diminish its impact on their daily activities and increase their quality of life. It will also greatly reduce the economic burden of PD of health-care systems. Unfortunately, there is not a gold standard test in the diagnosis of this disease to clearly determine if a patient suffers from this condition. Diagnosis is based on the assessment of a patient's signs and symptoms over an observation period which can last from months to years.

Multiple studies have found enough evidences to propose biomarkers or automatic diagnosis schemes based on a patient's eye move-

ments of the patients [2], typing [3], drawing [4], hand-writting [5] or sign language movements [6] among others. However, in the clinical practice, none of the aforementioned approaches has materialized yet a standard test. Notwithstanding, the new machine learning technologies can boost the development of successful and reliable diagnosis supporting tools by demonstrating a reduction of the diagnosis time and the costs of the assessment process.

To this respect, voice and speech articulation are considered good candidates for PD detection but have not been deeply exploited yet for these purposes. Supporting diagnosis tools based on speech analysis are non-invasive, easy-to-use and low-cost solutions, being used for decades by specialists in clinical environments but not specifically to diagnose PD in an objective manner. However, most of the assessment procedures include perceptual assessments of the voice and speech of patients in addition to surveys assessing their communication abilities. The introduction of new methods of automatic objective assessment employing voice and speech signals can outperform the results of the perceptual assessments and therefore can be used to detect the signs that are not perceived in the first stages of the disease, but which could provide relevant information.

All of these facts have motivated the work included in this thesis, aimed to push the boundaries of the knowledge in the usefulness of voice and articulation in the objective assessment of PD to support the diagnosis.

1.2 INTRODUCTION TO THE DOCUMENT

In this document, there are four main parts, listed and described in the following lines:

1. Introduction: this part includes an approximation to PD, its causes and its consequences followed by the hypothesis and goals of the thesis, once the problem is explained. Then, the state-of-the-art of PD detection and assessment using voice and speech is analyzed in order to identify the different techniques applied so far, their advantages and their weaknesses as well as the unexplored regions of the knowledge to this respect. This part ends with an overview of the proposed methodology introducing the experiments to the reader.
2. Articulatory approaches: this part, composed of four chapters, describes the four experiments using articulatory information from speech to detect PD.
3. Phonatory approach: in this third part, an experiment analyzing the phonatory aspects of patients' voices is carried out to complement the previous experiments with a potential complementary approach.
4. Discussion, future work and conclusions: lastly, the results of all the experiments are compared and analyzed as a whole in the discussion. Future work and conclusions end the document.

All chapters describing experiments contain: an introduction to the problem; a theoretical background; an explanation of the methodology; results and a discussion; and a summary of the chapter.

PARKINSON'S DISEASE

In this chapter, an introduction to PD from different points of view is included to provide the reader with an overview of this condition and the reasons to seek for new schemes to help clinicians in the diagnosis and assessment.

There exist several types of diseases which are categorized as parkinsonisms. Among all of them, the most common is called idiopathic PD. In this thesis, all the references to PD are related to idiopathic PD unless specified. Most of the works in literature are focused in this type of parkinsonism when they use the term PD although many of them do not specify it.

2.1 DESCRIPTION AND PATHOPHYSIOLOGY

PD is a chronic and progressive condition caused by the gradual death of brain cells, including those located in the *substantia nigra*, implicated in the production of dopamine neurotransmitters. This neurotransmitter is involved in many neuronal activities playing an important role in motor tasks. The consequent loss of dopamine results in a lack of coordination, muscle rigidity and slowness of movements, among other symptoms, in the subject affected by PD. It was first described in 1817 by James Parkinson in his work *An essay on the shaking palsy* [7]. Although the disease was described then, it is not until the second half of the twentieth century that the loss of dopaminergic cells was identified as the main cause of the symptoms in PD.

2.2 ETIOLOGY

The causes originating idiopathic PD are still unknown. In fact, the term *idiopathic* comes from Greek *idios*, that means "one's own" and *pathos*, "suffering" or "disease", and is employed with diseases of uncertain causes. However, two main factors that have been revealed as influential in developing the disease are:

1. **Genetic causes.** A percentage of the patients with this disease (usually around 10% although in some ethnic minorities it can reach 40%) have a genetic mutation leading them to develop PD. The Leucine-Rich Repeat Kinase 2 (*LRRK2*) [8] is the most recognized gene associated with increasing an individual's risk for developing PD. Although the symptoms of these patients can be the same as those of other patients, this type of PD is not considered strictly as idiopathic since its causes are known.

2. **Environmental causes.** Several studies link the disease to the herbicide and pesticide exposures and other toxic substances present in the drinking water or directly suspended in the air. For this reason, some studies show a higher prevalence in rural and farming environments [9], [10]. However, these are not the only environmental factors. Literature also suggests that the risk of suffering from PD is higher in subjects with melanoma, or those who have undergone a Traumatic Brain Injury (TBI). The literature also suggest an increased risk linked with a high consumption of dairy products and methamphetamine, while it is decreased in subjects with high blood cholesterol and hypertension. Protective factors include regular tobacco smoking, coffee, green and black tea, antioxidants such as urate, and physical activity [11], [12].

In addition to these two causes, **age** is the main known risk factor, being more likely to find subjects affected by PD in populations older than 60 years.

The work [12] suggests that the environmental factors have more or less impact in subjects depending on their genetic causal or protective factors. Figure 2.1 represents a diagram of all of these factors influencing the PD occurrence.

Figure 2.1: The balance of causal and protective environmental factors weighted by the causal and protective genetic factors in the PD occurrence. Extracted from [12]

2.3 SYMPTOMATOLOGY

PD entails very heterogeneous symptoms, from which many of them are shared with other conditions. This is one of the reasons why a precise diagnosis is complicated and requires long periods of observation. The most common signs are related to movement although there are a

variety of non-motor signs like smell loss or constipation. Among the motor symptoms, the most common are:

1. Akinesia: a delay in the beginning of movements with long reaction times.
2. Bradykinesia: slowness of movements.
3. Hipokinesia: poor, incomplete or simplified movements. In most cases, the limbs involved do not reach the full extent associated with the action of purpose of the movement. As an example, handwriting can become smaller.
4. Rigidity: rigidity of muscles, causing pain and hampering certain actions and postures.
5. Tremor: it can be postural or resting tremor. In the same manner, it can be constant or intermittent. In some cases, it consists of an internal tremor, not visible but quite disturbing for the patient.
6. Postural instability: involving balance problems in the standing position or walking.

Regarding the non-motor symptoms, the most common are:

1. Depression and anxiety. It can involve the inability of the patient to express emotions.
2. Constipation.
3. Loss of smell.
4. Swallowing problems.
5. Communication problems.
6. Dementia and other cognitive problems such as hallucinations or difficulties on focusing and performing complex tasks.
7. Sleeping problems.
8. Sexual disruptions.

Usually, the motor symptoms are more or less evident since the onset of the disease, but these are different depending on the patient. And this is a characteristic of [PD](#): each parkinsonian patient experiences the symptoms (and response to treatment) in a different manner. Unlike other conditions, there is not a clear pattern, which leads to a more complex and uncertain diagnosis. On the other hand, some studies evidence that other non-motor symptoms such as constipation or sleeping problems are manifested up to 20 years before the disease is diagnosed [[13](#)], which opens the door to the existence of early biomarkers which, eventually, would lead to an early diagnosis.

2.4 EPIDEMIOLOGY AND PREVALENCE

Recent studies [14] point out that, in industrialized countries, PD affects to 0.3% of general population to 3% of olders than 80 years and to 1% of subjects older than 60 years, being this last one the median onset age. Men are more likely to develop PD, with a risk from 1.5 to 2.0 times higher than women [15]. As it is stated in [1], in 2005 more than 4.1 millions of people suffered PD in Western Europe's five most and the world's ten most populous nations. The same study estimates that based on several facts as the growth of life expectancy, the number of PD patients could reach 9.3 millions by 2030 in these countries.

2.5 DIAGNOSIS AND RATING SCALES

As it was said in the introduction of this document, an early detection of the disorder might improve and maintain the quality of life of the patients and increase their life expectancy [16] but unfortunately there are not efficient and reliable methods capable of achieving this early diagnosis.

The most common criterion for PD diagnosis are mainly based on the observation of motor *cardinal signs* [17] and other non-motor indicators as dementia, depression, excessive salivation, constipation in addition to other physiological and cognitive manifestations which are employed in the so called *clinical diagnosis*. Notwithstanding, neuropathological diagnosis during autopsy is considered as the gold standard although some studies demonstrate that following the usual clinical diagnosis criteria it is possible to obtain 90% of accuracy in final judgement, being 2.9 years the average time to obtain it [18].

Following the diagnosis procedure proposed in [19] there are three different levels of diagnostic confidence: possible, probable and definite. The two first are based on clinical criteria and the last requires neuropathological confirmation. In this work, the clinical features are sorted into cardinal signs, asymmetric onset, levodopa responsiveness, other motor features and nonmotor clinical features:

1. Cardinal signs: are the four most typical signs of PD and consist in tremor, rigidity, bradykinesia, and postural instability. Tremor, or more commonly resting tremor, is basically a trembling below 6 Hz affecting to, around 70% of PD patients [20]–[22]. Rigidity is characterized by the high resistance of the muscles to passive manipulation, thus the clinical testing involves passive manipulation of arms and legs. It appears in a range from 89% to 99% of the patients [20], [21], [23]. Bradykinesia, or slowness of movements entails the difficulty of initiating a movement. Other similar movement perturbations are akinesia (absence of movements) and hypokinesia (poverty of movements). It appears in almost all patients and it is crucial for diagnosis although it is not ex-

clusive of PD as it can be found in several degenerative disorders as other parkinsonian diseases or Alzheimer. Regarding postural instability, PD patients have difficulty adapting their posture for changes in postural demand [24], as for a variation in the environment requiring the readaption of the body position. This instability does not occur until advanced stages and also appears in other neurodegenerative disorders, having, therefore, limited diagnostic specificity.

2. Asymmetric onset: in more than 70% of the cases, the symptoms start asymmetrically [21] and although PD is not the only disease in which the symptoms appear asymmetrically, it is a characteristic manifestation of PD.
3. Treatment responsiveness: most of the PD patients (from 94% to 100%) respond favorably to levodopa [19] although this is not exclusive of these patients.
4. Other motor features: speech and blink disturbances, festination, and other motor signs.
5. Nonmotor clinical features: there are other characteristics which are observable in a lower portion of the patients, usually less than 40% in the first stages, such as dementia, depression, olfactory dysfunction and oculomotor abnormalities among others.

The existence of these signs provides information about the confidence of the diagnosis but none of them is definitive or has sufficient specificity to detect PD as other neurodegenerative diseases can present the same signs. The conclusions extracted by Gelb et al. to propose a clinical diagnostic criteria are that the combination of clinical features can improve accuracy, taking into account that some of them are more useful than others. Additionally, the probability of missing diagnosis is higher in the early stages so the more probable diagnosis must wait until the symptoms stay during a long period and there are no other signs suggesting the possibility of other disease. At this respect, Gelb says: *"Patient diagnoses are classified as Probable when the patients have clinical features that are absolutely typical of PD, at least 3 years of parkinsonian symptoms without the development of atypical features, and a clear clinical response to dopaminergic treatment."* Tables 2.1 and 2.2, extracted from [19], list a common categorization of clinical features and diagnostic criteria of PD, respectively.

In the study [18] the accuracy and sensitivity of clinical diagnosis of 143 patients were studied during a decade. In the results, the probable diagnosis of idiopathic PD takes 1.4 years in average while in other parkinsonian diseases take 4.1 years. Moreover, the accuracy of clinical diagnosis is quite high for idiopathic PD, reaching up to 98.6%, while in other syndromes this value was 71.4%.

After the diagnosis and in order to monitor the progress of the disease, most specialists use the Unified Parkinson's Disease Rating

Table 2.1: Grouping of Clinical Features According to Diagnostic Utility. Extracted from [19]

Group A features: characteristic of PD
Resting tremor
Bradykinesia
Rigidity
Asymmetric onset
Group B features: suggestive of alternative diagnoses
Features unusual early in the clinical course
Prominent postural instability in the first 3 years after symptom onset
Freezing phenomena in the first 3 years
Hallucinations unrelated to medications in the first 3 years
Dementia preceding motor symptoms or in the first year
Supranuclear gaze palsy (other than restriction of upward gaze) or slowing of vertical saccades
Severe, symptomatic dysautonomia unrelated to medications
Documentation of a condition known to produce parkinsonism and plausibly connected to the patient's symptoms (such as suitably located focal brain lesions or neuroleptic use within the past 6 months)

Table 2.2: Proposed Diagnostic Criteria for PD. Extracted from [19].

Criteria for POSSIBLE diagnosis of Parkinson disease:
At least 2 of the 4 features in Group A are present; at least 1 of these is tremor or bradykinesia
and
Either None of the features in Group B is present Or Symptoms have been present for less than 3 years, and none of the features in Group B* is present to date
and
Either Substantial and sustained response to levodopa or a dopamine agonist has been documented Or Patient has not had an adequate trial of levodopa or dopamine agonist
Criteria for PROBABLE diagnosis of Parkinson disease:
At least 3 of the 4 features in Group A are present
and
None of the features in Group B is present (note: symptom duration of at least 3 years is necessary to meet this requirement)
and
Substantial and sustained response to levodopa or a dopamine agonist has been documented
Criteria for DEFINITE diagnosis of Parkinson disease:
All criteria for POSSIBLE Parkinson disease are met
and
Histopathologic confirmation of the diagnosis is obtained at autopsy

Scale (UPDRS) [25], which comprises four main parts: I, mentation, behaviour and mood; II, activities of daily living; III motor; IV complications. The rating according to this scale is accomplished through interviews with the patient and through clinical observations. The best possible score for each part is 0 whereas the worst one depends on the part: 16 for part I, 52 for part II, 56 for part III and 23 for part IV. Therefore, the global UPDRS can range between 0 and 147, where it can be claimed that, the larger the value, the higher the affection of the disease. Nevertheless, two patients with the same global rate can exhibit different signs.

A simpler scale widely used is the Hoehn & Yahr (H-Y) scale, comprising several stages whose values can range from 1 to 5 in which 1 implies that the patient has low or no functional disabilities and 5 that patient is totally dependent [20]. This scale was modified in [26] to include two intermediate levels (1.5 and 2.5) as listed:

1. Stage 1.0: symptoms are very mild; motor signs are usually unilateral.
2. Stage 1.5: unilateral and axial motor involvement.
3. Stage 2: bilateral involvement without balance impairments.
4. Stage 2.5: mild bilateral disease with recovery on pull test.
5. Stage 3: mild to moderate bilateral disease; some postural instability; patient is still physically independent.
6. Stage 4: severe disability; still able to walk or stand unassisted.
7. Stage 5: patient totally dependent. Wheelchair bound or bedridden unless aided.

At the same time, as authors in [27] say: "*Clinical diagnosis of Parkinson's syndrome is reasonably easy in most cases but the distinction between different variants may be difficult in early cases*", indicating that one of the main challenges is not only to detect a parkinsonism but to differentiate between the possible variants.

2.6 TREATMENT

The most common treatment for PD patients is Levodopa. This drug, is absorbed by the small intestine and transported to the central nervous system, where it is transformed into dopamine within the dopaminergic cells, counteracting the effects of PD. Normally, after 5 years of pharmacological therapy (usually, Levodopa), at least 50% of the patients, start to experience the "ON-OFF" phenomenon, which modulates the presence and intensity of the symptoms [28]. The patients suffering from this phenomenon undergo through different periods within a day in which some times their symptoms are alleviated (ON

phase), and others in which the patient responds poorly to the medication (OFF phase). Commonly, the ON phases occur right after the Levodopa intake and the OFF phases, before the next dose. As the disease evolves, the "on" phases tend to be shorter and the "off" phases, longer. As a patient points out in [29]: *"ON is quite simply normal; I can survive a dinner party, drive a car, write a fair, round hand, my voice is normal. I can fall asleep rather easily unless I am trying not to. OFF on the other hand is very unpleasant. I lose almost all motor power in my legs; and this paralysis increasingly now spreads to my arms. Sometimes odd pains and cramps move round the body. There is no position in which I am comfortable. I can't write, I can't type, my speech is slurred and low-powered."*

For this reason, recent studies [30] recommend other treatments such as monoamine oxidase type B inhibitors or dopamine agonists, among others, during the early stages, in order to avoid the Levodopa side effects and to extend the effectiveness of the overall treatment.

Within the group of neurological surgery treatments, the most common is the Deep Brain Stimulation (DBS), consisting in the implantation of electrodes in the brain generating electric impulses that regulate the abnormal impulses generating tremor.

On December 20st 2017, the company Denali Therapeutics announced very positive results on the first in-human [LRRK2](#) inhibitor clinical trial, proving that the experimental treatment is safe and effective in the inhibition of the [LRRK2](#) protein activity in human body cells. That means that this treatment could radically slow or even stop the advance of the disease progression in patients with genetic inclination to suffer from [PD](#). When this and other new treatments are discovered and introduced in the clinical practice, the early diagnosis will be crucial.

2.7 PARKINSON'S DISEASE AND SPEECH

It is well known that the neurodegenerative processes associated to [PD](#) cause dysphonia and dysarthria, in particular *hypokinetic dysarthria* [6], [31]–[33]. Although it is common to find both terms separately in literature, some authors bind them together using the denomination "dysarthrophonia" [34]. Dysphonia can be defined as the incapacity of the subject to produce a normal vocal sound or the existence of perturbations in the functioning of their phonatory system, while dysarthria is more related to their problems with articulation while pronouncing words. More specifically, *hypokinetic dysarthria* is characterized by a reduction of loudness and articulation amplitude, slow speech rates combined with rushes of fast speech sometimes and a decrease of intelligibility mostly.

Some studies analyzing directly the movement excursions of the articulators indicate that [PD](#) patients exhibit decreased excursions and lower velocities of tongue [35] and lips [36] than neurological healthy subjects. The most accepted hypothesis is that the impairments in voice and speech are related to the rigidity and bradykinesia caused by [PD](#)

which produce uneven vocal folds vibration, incomplete folds closure and undershooting of articulatory gestures due to the decreased range of motion of the supra-laryngeal muscles as well as in the respiratory muscles [37]–[39]. Literature demonstrates the influence of PD in speech from early to advanced stages [34], [40] but hypokinetic dysarthria is mainly noticed in mid to advanced stages of the disease and can be used to quantify the level of affection [41], [42]. Notwithstanding, some works as [43] report cases in which family members and relatives of the patient perceived changes in oral expression even before the disease was diagnosed. The articulatory deviations in these preliminary stages could not be clearly perceived due to compensation mechanisms as those suggested in [44]. In other cases, patients are not self-aware of their voice/speech handicap [45] and the perception of the changes in their voice and speech is different than the one reported by their family and relatives. But in general, and depending on the stage of the disease, most of the PD patients have, at some point, speech communication difficulties, specially by phone, which eventually can lead to social withdrawal [46].

Some specialists suggest that 90% of PD patients after a median latency period of 7 years since diagnosis suffer from dysarthria [41], [47], which is in accordance with the work [48], although other works such as [49], [50] point to lower values. Recent works like [51], employing perceptual estimations or preliminary quality of voice analysis, have quantified to 100% the percentage of patients with signs of PD in their voice. To determine this value, authors have used the Robertson dysarthria profile [52] in a cohort of 48 patients in several stages. In the same sense, work [34] reports anomalies in at least one out of all the analyzed aspects of voice and speech of all the 80 patients of the study, independently of the stage of the disease. In this work, in contrast to most of the previously mentioned, not only a perceptual but an exhaustive acoustic analysis is performed. In any case, although the methodology followed is precise, the number of patients (specially those in an early stage) is relatively small to reach strong conclusions.

Authors of [50] perform a perceptual analysis of phonatory, articulatory and prosodic dysfunctions in a large cohort of speakers (200 idiopathic PD patients), reporting deterioration of voice and speech functions in more than 73% of the patients. Most of the patients present voice Sound Pressure Level (SPL) or voice intensity¹ reduction and phonatory problems, with added articulatory deficits or fluency impairments in some cases. Authors suggest that these fluency and articulatory deficits rarely exist in isolation.

---

¹ Although many works related to voice analysis use the expression *voice intensity*, in this thesis *voice SPL* or just *SPL* are used instead since it is considered more correct, as the word *intensity* in acoustics is referred to the power per unit area in a direction perpendicular to that area and what is usually measured to characterize the loudness of a voice is the *SPL*.

Regarding early stages, the main conclusion of work [40] is that 78% of the early untreated studied patients exhibit some form of vocal impairment. In this work, phonatory, prosodic and articulatory aspects are studied in 23 patients and 23 controls. This conclusion is highly relevant as it indicates that speech could play an important role in the early diagnosis of PD. In this sense, Sabine Skodda finds perceptual and objective deterioration of phonatory, articulatory and prosodic aspects of speech in all 80 patients of the study [34], independently of the stage of the disease, indicating that abnormalities on voice and speech can occur in early stages.

Therefore, it is difficult to conclude which percentage of patients has a reflection of PD in voice or speech since most of them have not studied in detail the perturbations in voice during the early stages neither the differences between the dysarthric disturbances caused by PD and aging separately and in the cases in which these considerations are made, the number of patients of the study cannot be considered strongly representative. Nevertheless, most of the works in literature agree that the speech impairments are more pronounced and noticeable as the disease progresses [34], [48], [50], [53].

The treatment of PD dysarthria usually involves the practice of several exercises under the guidance of a speech therapist as in the phonatory and respiratory effort (the Lee Silvermant Voice Treatment (LSVT)) which has reported great results in improving the voice and speech capabilities of the patients [54], [55]. Nevertheless, some treatments as the one proposed in [56] are aimed to counteract the effects of the disease by modifying the structure of the vocal folds. This technique consists in the collagen augmentation of the vocal folds of patients with glottal insufficiency, reporting, in this case, a 75% of patient satisfaction after a telephone survey.

Respecting the influence of pharmacological treatment on speech, an early work [57] describes the effects of levodopa on 18 patients and reports increased speed and enhanced symmetry and motility on the lips, providing an enhancement of intelligibility compared with the conditions before the treatment. A more recent work [58] performs several phonatory, articulatory and prosodic based- measurements on 19 patients before and after the beginning of antiparkinsonian therapy. Changes on articulatory and prosodic aspects after the beginning of the treatment and a correlation with bradykinesia and rigidity are described. Authors consider that, therefore, some of the measurements on speech could be used as a valuable indicator of the evolution and treatment efficacy on PD. This work is in agreement with the findings reported in several early studies like [59]–[61] although other more recent works illustrate no significant differences in most of the aspects of speech before and after a maintained pharmacological treatment [62]–[66]. Authors in [58] attribute these differences due to patient-related particularities and methodological differences. They argue that most of the previous works use simple speech tasks and focus in a restricted

number of features, a single speech aspect or speech assessment procedures as the [UPDRS](#) speech scores which seem to be insufficient to evaluate the changes in speech before and after treatment. For instance, [\[66\]](#) does not find any statistically significant differences on Relative Fundamental Frequency ([RFF](#)) between patients before and after medication intake but that does not mean that other features remain unchanged.

Other works have explored the impact on speech after some treatments such as the [LSVT](#) [\[55\]](#) or [DBS](#) [\[67\]](#). The latter treatment is aimed to reduce tremor, rigidity and bradykinesia by electrically stimulating the subthalamic nucleus and these reductions are expected to improve the patients' speech. However, only some works show a moderate improvement on the patients' communication abilities, specially in advanced stages [\[68\]](#), [\[69\]](#) while others report a negative effect [\[67\]](#), [\[70\]](#), [\[71\]](#). As the influence of this treatment on speech seems to depend on the patient, authors of [\[67\]](#) recommend to follow a methodology to analyze the speech signal after the implantation of [DBS](#) in order to choose the correct settings not to impair the patients' communication capabilities.

To summarize, although there is not consensus about the percentage of patients with anomalies in voice or speech caused by [PD](#), it is possible to say that at least 80% of patients suffer from any type of impairment related with the voice or speech. Pharmacological treatments have an influence on some aspects of speech which is not always evident while subthalamic electrical stimulation and speech therapies have a clearer manifestation in this sense. Additionally, despite of the presence of objective acoustic cues which link [PD](#) and speech, the majority of the clinical procedures using speech to assess [PD](#) consists in perceptual assessments [\[41\]](#). Thus, this is only valid when the hypokinetic dysarthria is perceived and, therefore, is not useful for an early detection. In this regard, it is expected that new methods of automatic assessment employing voice and speech signal can provide relevant information about the presence or evolution of [PD](#).

HYPOTHESIS AND GOALS

3.1 HYPOTHESIS

Phonatory and articulatory speech aspects of the patients with idiopathic PD are affected by the motor dysfunctions associated with this neurodegenerative disease. The resulting speech impairments can be characterized using signal processing techniques. These can be tremor, noise, frequency and amplitude perturbations in the glottal source and inaccuracy of articulatory movements as well as modified patterns of acceleration and velocity of the articulators. This characterization, combined with state-of-the-art classification schemes employed in speaker and speech recognition, can help in the differential diagnosis of PD.

3.2 GOALS AND OBJECTIVES

The main goal of this thesis is to analyze distinct advanced speech processing technologies and machine learning techniques for the detection and assessment of PD from articulatory and phonatory aspects of speech. The resulting systems will serve as supporting tools in the differential diagnosis of PD for new patients.

The specific objectives derived from the main goal are:

1. The analysis of supervised schemes to detect PD from phonatory aspects of speech.
2. The analysis of several supervised and unsupervised new schemes to detect PD from articulatory aspects of speech.
3. The analysis of the role of the distinct phonetic unities in the automatic detection of PD. Although the literature suggests the influence of PD in certain segments as the silences before the bursts in plosives, or directly plosives and fricatives, no work has thoroughly studied the influence of these segments in automatic detection systems.
4. The analysis of the distinct speech transitions between phonetic unities during articulation and their influence in the detection of PD.
5. The identification of the advantages and disadvantages of the use of different speech tasks in the proposed schemes to detect PD.
6. The analysis of the combination of a phonatory subsystem with an articulatory subsystem to automatically detect PD.

7. The study of articulators' kinematics and its relevance in the automatic detection of [PD](#).
8. The analysis of the correlation between the scores of the proposed systems and the severity of the disease.
9. The study of the generalization properties of the analyzed systems.

STATE OF THE ART

In this chapter, a review of the different methodologies related to the analysis and characterization of voice and speech in parkinsonian patients for the automatic detection of PD by using these signals is performed. The main signal processing technologies that have been explored so far to provide new biomarkers and assess the grade of affection of the disease are outlined too.

4.1 INTRODUCTION

Although there are several works previous to 1970 describing parkinsonian speech [72], [73], it is not until 1969 [31], [74] that a group of researchers analyzed in a more thorough manner the problems of phonation, prosody and articulation of PD patients.

Literature shows a certain number of studies and approaches trying to identify the deficiencies caused by PD in the voice or speech of patients or to propose the creation of new biomarkers or automatic detectors to support the diagnosis of this condition. In this thesis, the studied works are divided into four groups depending on the analyzed speech aspect: phonatory, articulatory, prosodic and linguistic. Each of these aspects, could be subdivided, at the same time, into further dimensions. For instance, phonatory aspects can be related to the phonation activity (source) or to the resonant cavities configuration. The prosodic aspect can be studied observing the speech rate, intonation, accentuation and regularity while the articulatory aspect's approaches focus on the segments length, transitions analysis or phonetics, among others. The linguistic works can be subdivided into those studying repetition of words, use of vocabulary or employment of certain expressions among others.

With the aim of introducing these aspects, it is possible to say that the phonatory studies are related to the glottal source and resonant structures of the vocal tract in an static manner, employing most of the times sustained vowels as acoustic material. A great number of them analyze phonatory perturbations utilizing acoustic measurements such as jitter, shimmer or noise, or try to characterize the supra-laryngeal structures by calculating formant-related features such as the Vowel Space Area (VSA).

Works based on articulatory aspects are more diverse as there exist more analysis possibilities. This is due to the fact that the obtention of features or the acoustic measurements can come from different types

of segments of phonemes¹, the velocity or acceleration of articulators, transitions between segments or evolution of formants among others. However, there are fewer number of articulatory studies in literature than, for instance, phonatory or prosodic due to the complexity of this type of analyses even when the influence of PD on articulatory processes seems to be clear [41].

Concerning the prosodic works, studies are mainly focused in the paralinguistic features such as pitch variation, syllable rate analysis, or the manifestation of emotions among others.

Finally, the linguistic approaches analyze the deviations in the cognitive behavior by examining the vocabulary, phrase construction and the existence of word repetitions.

The speech task or acoustic material used in each case is a differentiating key factor of the four groups. In the phonatory analysis, the most advisable acoustic material is sustained vowels while in the other three groups, running speech is needed. In cognitive studies, spontaneous speech, questionnaire answers and topic-based speech are almost unavoidable. For prosody analysis, fixed text including different modes (neutral, interrogative, imperative, etc.) and emotional speech are used in addition to spontaneous speech. Finally, in articulatory analysis, DDK tasks² can be valuable in addition to the other running speech materials such as spontaneous speech or fixed text.

Among all the works performing an evaluation of the patient's speech and voice it is possible to differentiate between *perceptual* and *objective assessments*, depending on the type of analysis. In the first type of assessments, trained evaluators following an specific protocol such as Grade, Roughness, Breathiness, Asthenia and Strain scale (GRBAS) [76] or Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) [77] are necessary. On the other side, the objective assessments make use of algorithms to characterize the signal through certain features. The studies framed within the prosodic and articulatory aspects, in comparison to the phonatory ones, are based on the feature extraction from running speech, or the processing of certain segments of the speech. Still, in many cases these features coincide with those employed to characterize phonatory aspects or contain information related to them

¹ A phoneme is a phonetic unit which cannot be decomposed in minor unities and that can distinguish one word from other. For a more detailed information about phonetics, go to section 7.2.

² DDK tasks consist in the repetition of words or syllables starting in a slow or calm syllable rate which is increased until the speakers reach their limit rate. Literature reports significant differences between controls and patients in several measures over DDK utterances [75]. DDK sequences are of great interest because these tasks have associated alternating articulatory movements where, usually, the employment of plosive syllables and different points of articulation promotes a good scenario in which changes in velocity and acceleration of articulators can be used to assess articulatory impairments. DDK analysis is normally performed in two different manners: when there is a repetition of a single syllable (alternating motion rates can be measured) and repetitions of a sequence of syllables (sequential motion rate).

such as fundamental frequency or formants frequency, but measured on running speech.

From those works found in literature performing an objective assessment, some propose measures or biomarkers for detection of PD or correlated with the stage of the disease. Others propose automatic binary classifiers to distinguish between parkinsonian speakers and controls or to automatically predict the stage of the disease. Attending to this last type of studies, it is possible to observe another possible categorization containing two principal groups: *supervised* and *unsupervised approaches*. The *supervised approaches* are those in which the system must be supervised at some point to perform an action (such as a manual segmentation of the speech or fragment selection) or there are constraints in the speech tasks to be used. These systems require that the speaker pronounce a specific sentence, a sustained vowel or a DDK task, usually under certain controlled conditions and are more suitable for clinical environments. Most of the works found in literature are of this type. The *unsupervised approaches* normally make use of free running speech and have not restrictions on the speech task to be used as the input of the detector. Moreover, in the *unsupervised approaches* no action from a supervisor is required. These systems could be used in both, clinical and domestic environments, and are suitable to monitor the evolution of the patient as they can be installed on devices sampling large speech sequences in a domestic environment. These last ones are still emerging approaches and it is possible to find only a few contributions in this sense in literature.

In the same way, the systems can be language dependent or non-dependent. Usually, phonatory approaches using sustained vowels are language-independent. Most of the works found in literature are language independent as none seems to use one specific allophone³ or type of articulatory movement related to one specific language. Notwithstanding, applying a methodology to different languages could provide different results. In the same sense, most of these studies train acoustic models with databases in a single language and, although the methodology could be applied to other languages, the obtained models are exclusive for the training language.

In this chapter, four main sections review the state of the art about the study of parkinsonian speech depending on each aspect (phonatory, articulatory, prosodic and linguistic), with a special attention to those focused in detection and assessment of PD. Within each of these sections, the other categories explained in this introduction (as supervised or unsupervised approach, or the subcategories within the aspect) are mentioned when corresponding. Additionally, a section referred to the combination of different aspects into a single approach,

---

³ The denomination "allophone" is referred to each of the pronunciation variants of a certain phoneme, usually, as a function of this phoneme in a word or syllable and depending on the adjacent phonemes

a review of some of the parkinsonian speech databases found and a discussion about all the exposed findings are included.

4.2 PHONATORY ASPECTS

4.2.1 *Evidences of the influence of PD in the phonatory system*

As it has been said in the previous section, the phonatory approaches related to [PD](#) mainly analyze the impairments in phonatory-related structures and muscles as the diaphragm, the muscles connected to the larynx, the vocal folds or the supra-glottal cavities and articulatory muscles of patients. The **type of acoustic material** analyzed in these cases is usually the voice signal from one or several **sustained vowels** which can be maintained during a certain lapse of time or modulated in amplitude or frequency, depending on the features to measure or assess. Sustained vowels are expected to generate simpler acoustic structures that might lead to consistent and reliable, to some extent, perceptual judgments of voice quality. Therefore, this perceptual judgment is employed in multiple studies analyzing the influence of [PD](#) on phonation, as it is detailed in this chapter. One of the possible reasons influencing the existence of a larger number of phonatory studies than of any of the other speech aspects of parkinsonian patients, is the fact that objective phonatory analysis is relatively easy to perform, as acoustic measures can be directly made using available software and libraries such as the Multidimensional Voice Program ([MDVP](#)) [[78](#)], Praat [[79](#)] or OpenSMILE [[80](#)].

In one of the earliest works, [[48](#)], authors **analyze perceptually** the dysfunctions of voice and speech on 200 untreated patients (idiopathic and postencephalitic). After assessing their quality of voice, the study reports laryngeal and/or vocal tract abnormalities in 89% of the patients, comprising hoarseness, breathiness, tremulousness and roughness while only a 10% exhibits hypernasality. In a subsequent work, [[81](#)], authors examine parkinsonian patients -32, in this case- by means of telescopic cinelaryngoscopy and report rigidity in the phonatory posture of the larynx and a high correlation with limbs rigidity. That means that if the rigidity found in some of the patients is higher on the left limbs, the same effect is found in the left vocal fold, compared to the right fold. This increased muscular activity produces in most of the cases a bowing of the vocal folds and, thus, a significant glottal gap during phonation which leads to weakness and breathiness. Although this effect cannot always be perceived, it is measurable. In concordance with these observations, work [[82](#)] demonstrates that it is possible to detect the influence of [PD](#) in the vocal folds vibration by reason of an excess of noise and other perturbations in the phonatory signal, caused by abnormal phase closure and phase asymmetry or vocal tremor resulting from abnormal vibrations visualized in the movements of the arytenoid cartilages. In a direct relationship with these conclusions,

the work [83] reports a significantly higher subglottal pressure in patients than in control subjects for the same phonation level, indicating that the mean laryngeal resistance of patients is greater too. This observation correlates with the patient's perception that they are working harder to produce phonation, as it is exposed in the same study. This can be caused by a greater resistance of the vocal folds to deformation, and therefore, patients need to perform a higher respiratory effort to compensate the increased resistance effect and produce the desired phonation level. Additionally, the authors of [83] suggest that, after comparing this resistance and the vocal flow of patients and controls, the former suffer from a significant incomplete glottal closure during phonation, leading to a breathy voice. Work [45] reports differences in the laryngeal resistance and the subglottal pressure of the two studied groups but, in discrepancy with [83] these values are lower in patients, although this could be explained by the fact that in this new study, controls report on average higher sound pressure level during phonation and in [83] patients and controls were asked to sustain a vowel producing in all cases similar levels.

Finally, and in line with previous findings, publication [84] describes, equally, an incomplete closure of vocal folds in patients after videolaryngostroboscopic examination.

4.2.2 *Acoustic measurements and perceptual evaluations*

One of the first attempts to perform **acoustic measurements** in patients can be found in [85] where jitter, shimmer and Harmonic to Noise Ratio (**HNR**) [86] are measured in the sustained phonation of /a/ from patients and controls, obtaining significant differences between groups for the three measurements, being jitter and shimmer higher and **HNR** lower in patients. Authors in [87] consider that jitter and shimmer should reflect the stability of the phonatory system and, thus, this frequency and amplitude perturbations along with **HNR** and the fundamental frequency are measured in sustained /a/ (sampling rate of 20 kHz⁴) of 22 untreated patients and 28 controls. The statistical differences found between groups suggest that **HNR** is lower and fundamental frequency is higher only in **PD** female patients and jitter and shimmer are higher in the whole parkinsonian set. In contrast, the work [88] finds a more prominent jitter only in patients in the later stages of the disease while non-significant statistical results are encountered for shimmer and Noise to Harmonics Ratio (**NHR**) using a sampling rate of 25 kHz. Accordingly, the study [89], using 20 kHz as sampling rate, reports significance of jitter and percent jitter on discriminating between patients and controls only in the male group and no significance for shimmer and percent shimmer. In [40] authors

⁴ When possible, sampling rates are reported in this text when analyzing jitter and shimmer results as these two features are sensitive to sampling rate.

employ sustained vowels to measure acoustic parameters such as jitter, shimmer, fundamental frequency, [NHR](#), [HNR](#) or Voice Onset Time ([VOT](#)) (which is more an articulatory feature) and results demonstrate that jitter, shimmer and noise are, in average, higher in patients respecting controls.

Other articles as [\[84\]](#) find differences in distinct acoustic measurements (specially in maximum phonation time) and perceptual ratings between patients of both sexes and their respective controls. However, in this case, no significant differences were found between groups in jitter and shimmer. The study [\[89\]](#), with 41 patients and 40 controls, performs nonlinear dynamic analysis, measuring Correlation Dimension (D_2), which turns out to provide significantly higher values in female patients than in female controls. More recent works like [\[90\]](#) use the [MDVP](#) to measure 24 acoustic features including frequency and amplitude perturbation, noise and tremor among others in 39 patients and 62 controls. Amplitude and frequency perturbations (shimmer and jitter families, respectively) result to be higher in patients of both sexes, and authors attribute this behavior to a reduced short-term neuromuscular control of the laryngeal abductory or adductory mechanisms. Higher values of [NHR](#), Soft Phonation Index ([SPI](#)) and Fundamental frequency tremor ([Fftr](#)) are found in patients respect to controls possibly caused by the muscular rigidity and tremor present in the larynx.

To this respect, authors of work [\[91\]](#) claim that sustained vowels seem to provide better results than other types of materials in distinguishing between parkinsonian and control voices. However, in this work, most of the used features are typical phonatory measures such as jitter, shimmer, [HNR](#), [NHR](#) or pitch statistics obtained employing Praat software to analyze sustained vowels, but words and short sentences too. Thus, these methods are not enough to state that sustained vowels (and, therefore, phonatory analysis) provide better results in the detection of [PD](#) as most of the used features are aimed to characterize mainly phonation, not articulatory characteristics while materials including articulatory information are used (words and sentences). The use of jitter and shimmer features in running speech could be considered unorthodox, specially when it seems that authors have used the complete signal for the analysis (authors do not specify if they use only vocalic segments from running speech).

Sabine Skodda performs a perceptual assessment as well as jitter, shimmer and [NHR](#) measures over sustained vowels in a longitudinal study [\[34\]](#) in which several speech tasks are recorded twice in 80 patients (with a mean time between the examinations of 33 months) and 60 age-matched controls (with a mean time between examinations of 25 months.) A significant correlation is found between the perceptual assessment of voice and the [H-Y](#) stage along time. Higher values of shimmer and [NHR](#) are found in male and female patients regarding their respective control groups while jitter does not seem to provide differentiation between them.

Another example of a longitudinal study can be found in [58] where Jan Rusz analyzes phonatory, articulatory and prosodic aspects of 19 Czech patients before and after being treated with symptomatic medication (mostly, levodopa or a dopamine agonist). Regarding the phonatory analysis, jitter, shimmer, HNR, Pitch Period Entropy (PPE)⁵, Recurrence Period Density Entropy (RPDE) and Maximum Phonation Time (MPT) are measured, finding statistical significant differences between patients and controls in all of these measures except by MPT and RPDE. These results are found before and after the beginning of the pharmacological therapy although the quality of voice seems to improve after the treatment, attending to the changes in jitter, shimmer, HNR, PPE and RPDE.

In general, it is possible to find some works like [87], [89], [92], [93] that find **differences** in the impact of PD over phonation **for male and female**. To this respect, in [92] authors identify distinct alterations in spectral energy distribution as well as abrupt fundamental frequency shifts in female patients respect males and controls. They conclude that this can be explained by the differences in the shape and size of the larynx of men and women, which can explain, in this case, the differences in vocal dysfunctions.

4.2.3 Formant measurements and associated features

Certain studies in literature analyze formant-related features, as the VSA obtained from sustained vowels, to separate patients from control speakers. An example of these works is [94], where authors measure jitter, shimmer, NHR and the triangular Vowel Space Area (tVSA). From this study it can be deduced that patients exhibit a higher jitter and NHR as well as a decreased tVSA, regarding to controls. One of the main drawbacks of this work is that authors only consider 9 control speakers and 9 patients, most of them in an mild to advanced stage of PD.

However, in [95], a comparison of PD detection techniques is performed using only vowel segments extracted from sustained vowels, sentence repetitions, reading passages and monologues of 20 untreated patients and 15 male control speakers. Significant differences between the two groups are obtained for tVSA and Vowel Articulation Index (VAI) [96], features which are lower in patients than in controls. One of the main divergences between [94] and this work is that in the second one, authors compare the use of vowels extracted from running speech and sustained vowels, obtaining enhanced results and more significant separation between groups when using vowels from monologues.

⁵ The functioning of this feature is explained in section 4.2.4.

4.2.4 *Automatic detection and assessment*

Within all the works in literature addressing phonatory aspects, some of the most recent use certain features, specific or already used in other studies about voice pathology detection, as the input of automatic detectors to distinguish between parkinsonian and normophonic voices or to predict the stage of a patient respect to the disease by means of machine learning technologies. The following works are framed within this group, which require a special analysis as this thesis could be included in it.

In this sense, some works like [97]–[99] use dysphonia measures to assess the severity of PD in telemonitoring scenarios achieving promising results. In the first one, authors perform a wide number of typical acoustic measurements in 31 speakers (23 with PD) including, among others, jitter, shimmer, HNR, NHR using the MDVP and other complexity measurements such as D_2 and a new feature proposed by them called PPE to discriminate between parkinsonian and normophonic voices. The proposed PPE is included in the study as a new feature aimed to quantify the abnormal variations of pitch, measuring the entropy of the probability distribution of the variations of the pitch in logarithmic scale (semitons) excluding the smooth vibrato and microtremor, commonly present in normophonic speakers. Some of the features are used in a Support Vector Machine (SVM) classification system after a previous feature selection to remove redundant information, in order to determine which pairs of features have a better discrimination in the SVM multispace. Results reveal that complexity (RPDE, Detrended Fluctuation Analysis (DFA) and PPE) and noise (HNR) measurements may have better discriminative properties, but the study only provides graphical results. In [98], authors use almost the same acoustic measurements and their log-transformation to predict UPDRS through linear regression in a longitudinal study in which the materials consist on almost 6.000 sustained phonations recorded weekly from 52 patients during 6 months as described in [100]. The mean absolute error of the predicted values for motor UPDRS is below 6.6, which is a good result considering that the rating can range between 0 and 108. The problem of this study is that in cross-validation (k-folds) authors select randomly the utterances which became part of the training and testing folds, meaning that both subgroups can contain utterances from the same speaker. To this respect, a mean error of 7.7 is obtained in a similar work, [42], using sustained vowels from 168 patients and features obtained with openSMILE in a more adequate cross-validation strategy. The same methodological issue can be found in the crossvalidation of the third work, [99], where 132 dysphonic measures are used, including the ones used in previous works and some additional such as Mel-Frequency Cepstral Coefficients (MFCC) or empirical mode decomposition excitation ratio, among others. A single vector containing each of these 132 measures is calculated for each of the 263 phonations

from the 43 subjects (10 controls and 33 patients) with a slight unbalance in age as the mean age of patients is more than 7 years higher than the mean age of controls, being the standard deviation higher too. Several methodologies of feature selection such as Least Absolute Shrinkage and Selection Operator (**LASSO**) or minimum Redundancy Maximum Relevance (**mRMR**) are applied following a crossvalidation scheme. The selected features are employed to train and test binary classifiers, which are Random Forest and **SVM** following a *k-folds* cross-validation scheme with 10 folds. In this case, an accuracy of 97.7% is reached in the detection of parkinsonian voices vs controls using 132 features and **SVM** with a radial basis function kernel and a 98.6% with the same classifier and only 10 features obtained after feature selection, consisting on 8 **MFCC** values and two noise features. However, as it is said, the use of the same speakers in training and testing partitions during cross-validation tends to provide over-optimistic results.

Other works, such as [101] perform an analysis of techniques to estimate 10 **PD** scores, including **UPDRS**, and to detect **PD** employing a corpus with 84 patients and 49 controls. Five vowels (/a/, /e/, /i/, /o/ and /u/) at minimum, normal and maximum **SPL** were characterized employing 350 features including jitter, shimmer, noise, complexity or empirical mode decomposition features among others. After using feature selection techniques and random forests for binary detection, an 90% accuracy is obtained employing all the vowels at minimum **SPL** as acoustic material (Leave One Out (**LOO**) cross-validation). The maximal error achieved in the estimation of scores does not exceeds 13%.

In a different approach, authors of [102] use Perceptual Linear Prediction (**PLP**) vectors to characterize sustained vowels of 34 subjects (17 patients). These features are the input of an automatic detector system (**SVM**) which provides a mean accuracy of 76% in discerning patients from controls. In a posterior work [103], same authors obtain an accuracy of 91% in a similar approach using **MFCC** with the same purposes, and in [104], authors report a 100% of accuracy training the system with a database containing 20 patients and 20 controls and testing it with another database containing 28 patients. However, these last results could be considered over-optimistic as there were no controls in this second database to test the discrimination properties of the system.

4.3 ARTICULATORY ASPECTS

4.3.1 *Evidences of the influence of PD in phoneme articulation*

Several works in literature point out to an evident influence of **PD** in phoneme production in patients, especially in the case of some specific allophones. One of the earliest studies trying to determine the articulatory deficits of parkinsonian patients from a phonetic point of view is [48] in which authors analyze the dysfunctions of larynx, lips and

tongue (back, tip and blades) in 200 idiopathic and postencephalitic untreated patients. In this work, two trained listeners evaluate the speech of the patients to perceptually assess their quality of voice and the misarticulation of allophones. Results show that a 45% of the patients exhibit lingual or labial (or both) abnormalities during articulation. Authors report that the errors are mainly concentrated in obstruent consonants, specially stop-plosives, fricatives and affricates (in this order), that is, the consonants requiring the greatest narrowing or closure during articulation, with more errors found in velar articulations, mostly /k/ and /g/ phonemes (this has been confirmed in subsequent works [105]–[108]). Additionally, the study points out that all of the patients with articulatory impairments have phonatory dysfunctions but not in the contrary. In a subsequent work [105] using the same database, authors report incomplete contact of articulators for plosive stops and partial constrictions for fricatives. One of the main conclusions found in this publication is that the intra-speaker consistency of the errors is near to 98%, meaning that when a patient misarticulates a phoneme, this error is repeated all along the session. However, this work does not include inter-session recordings and, thus, the longitudinal intra-speaker error consistency cannot be assessed. Another important observation found in this Logemann’s work is that inter-speaker type of error consistency reaches the 97%, meaning that the vast majority of the patients produce the same error substitution when a phoneme is misarticulated. In most of the cases, this error consisted in the substitution of stop phonemes by fricatives, in a phenomenon known as spirantization.

In relation with this study, the work [6] analyzes a DDK task (/pa-ta-ka/) of 16 patients and 10 controls, finding remarkable differences between groups in the duration of the voicing segments, which tend to be longer in patients and the silences before bursts or *voice bars*⁶ tend to disappear. These observations are in direct relation with the misarticulation of phonemes by which the stops are substituted by fricatives, as fricatives are not preceded by *voice bars*.

All the same, the work [110] presents a study of four speakers (two patients and two controls) along 10 years. A lower fundamental frequency variability is detected in patients, even before clinical diagnosis while VOT tends to be longer in patients in a middle stage. These results are confirmed in a second experiment detailed in the same publication and including four newly diagnosed patients.

One possible cause of the misarticulation can be related with disturbances in perception of the duration of occlusion lengths in allophones, as reported in [111]. Results of this study suggest that patients are inclined to perceive occlusion lengths longer than they really are, causing errors of identification of allophones. As an example, in the

⁶ The silences before bursts, or *voice bars*, as described in [109] are the pauses or silences before a burst on speech caused by a total constriction or closure in the vocal tract, specially during the articulation of plosives

cited work, some of the participant patients interpret the German word "boten" (/bo:tn/) as "boden" (/bo:tʰən/). Therefore, one hypothesis is that these perception impairments can contribute to the misarticulation during speech production as patients are not perceiving their own articulatory errors. The other hypothesis is that the mechanisms related to speech production and perception are related to cortical areas affected by the disease.

Kris Tjaden performs an analysis of the influence of two neurological conditions (PD and multiple sclerosis) on vowels [112] and consonants [113], asking speakers to read 25 sentences in habitual, clear, loud and slow conditions. The first work, using formant-related measurements such as VSA provides significant differences between the three analyzed groups (two groups of patients and one group of controls), in concordance with findings in other works [34], [55], [95], [96], [114], [115]. One of the conclusions of this work is that the production of vowel /u/ during articulation seems to be more affected by parkinsonism than /a/ or /i/. However, the second study, using features characterizing the articulatory constriction during consonants (spectral moment coefficients [116]) only suggests differences in the stops on medial position within words between controls and multiple sclerosis patients. The rest of the differences between the same consonants produced by the parkinsonian and control groups are subtle. This is contradictory with the findings exposed in [48], [105]–[108] where a strong influence of the disease is found in certain consonants.

It is important to take into account the complexity of the tasks when measuring articulatory features. The study [117] analyzes segments duration and speech error rates in 15 patients with neurological disorders (8 of them with PD) and 15 controls during the production of syllable sequences, ranging from low to high complexity⁷. Results suggest that patients present significantly longer vowel duration but the effects caused by the higher complexity of the sequences, as an increase of the speech error rate and a longer duration of speech measures, are similar in the two groups. Authors suggest that "the effects of phonemic similarity and phonotactic probability are primarily attributable to processing by distinct regions of cortex."

4.3.2 Kinematic considerations

Viewing these articulatory deficits from other point of view, some authors contemplate the kinematic properties of the articulators (mainly related to velocity and acceleration of these) as sources of information about the presence or evolution of the disease.

⁷ In the experiments carried out in [117], the highest complexity speech tasks are these syllabic sequences composed by several syllables including phonemically similar consonants and low phonotactic probabilities (/ja cha va za tha/) in contrast to the low complexity utterances (/ma ka/)

In the work [118], the author analyzes the dysarthria of 12 patients, comparing them with 10 controls performing a labial kinematics analysis at diverse speech rates. Results show significant differences between patients and controls in amplitude and velocity of lips movements at high speech rate (5 to 7 syllables/s) while no significant differences were found at low speech rates (3 to 5 syllables/s), suggesting that hypokinetic dysarthria varies as a function of the speaking rate.

In relation with amplitude and velocity, a work from the same year [36] reports lower amplitude and velocity in jaw and lower lip opening articulatory gestures for patients in comparison to controls (same conclusions are obtained in other works such as [35], [61], [118], [119]). In the same way, authors find reduced formant transitions and duration of vocalic segments. In concordance with this study, the work [120] shows a significantly reduced second formant (F_2) slope for patients (20 subjects) respect to controls (5 subjects).

Finally, in the approach [109], authors use the velocity and acceleration of the amplitude envelope of the speech signal during a DDK test to obtain several complexity measurements to discriminate patients from controls, with an accuracy around 85%.

4.3.3 *Formant measurements and associated features*

Several works employ approaches similar to [95], in which $tVSA$ is calculated obtaining the formants from vowels extracted from running speech, as described in subsection 4.2.3. Running speech contains fluctuations of vocal characteristics in relation to voice onset, voice termination, and voice breaks, which are considered to be crucial in quality of voice evaluation. These characteristics are not fully represented in signals such as a sustained vowel. For example, some studies perform the tracking of vowel formants and $tVSA$ with very similar results in their conclusions respect to this feature and the use of formants [55], [114], [115]. As formants reflect in some way the position of the tongue, a reduction of the articulation extension can subsequently limit the frequency range of the formants. For instance, in [55], authors record 14 patients before and after starting the LSVT, in addition to other 15 patients without this type of treatment and 14 control speakers. These last two groups are recorded twice too, with a similar time window between the two sessions in all the three groups. Three read sentences are recorded from each speaker at each session from where three vowels, /i/, /u/ and /a/ are extracted for analysis. Results indicate that the $tVSA$ is smaller in patients (in concordance with [93], [94]) and that it increases towards the control values after the LSVT. Nevertheless, same authors propose an alternative type of measurement based on formants, the Formant Centralization Ratio (FCR) [121], a metric being not sensitive to gender and more effective than $tVSA$ in distinguishing PD dysarthric speakers and controls. In a very similar way, the study [93] (68 patients and 32 controls) reports lower $tVSA$ in patients, spe-

cially for males, although results suggest that the use of the **VAI**, using the formants of the three vowels, /i/, /u/ and /a/, provides a better differentiation between parkinsonian and control speakers. From the same work it can be inferred that **tVSA** could be insensitive to mild forms of dysarthria. Very similar results are obtained in a subsequent work [114] (67 patients and 40 healthy speakers) where, additionally, **tVSA** and **VAI** exhibit a high correlation with gait impairment.

4.3.4 Automatic detection and assessment

Recently, new works propose automatic systems to detect or assess **PD** making use of speech processing advanced techniques. Going into detail about some of these studies, an analysis of **DDK** tasks (/pa-ta-ka/) in a database of 24 **PD** patients and 22 controls is performed in [122]. In this study all the utterances are subdivided automatically into syllabic segments and then, the initial burst, onset and occlusion positions are detected to analyze articulation. Only 13 features are obtained by performing measurements on the distinct syllabic segments, each feature describing a distinct articulatory trait of speech, such as the coordination of laryngeal and tongue activity or the precision of consonant articulation among others. Manual and automatic segmentations are tested obtaining up to 88% of efficiency in differentiating **PD** from control subjects. This study indicates that speech processing can produce powerful indicators of imprecise consonant articulation in **PD**-related dysarthria. Other works such as [123] use cepstral and frequency features like **MFCC** and Band Bark Energies (**BBE**) from running speech and features obtained after the segmentation of the speech to employ only unvoiced consonants, yielding good results with three corpora (in general, more than 80% of efficiency). This seems to point out to the hypothesis that there is more information about **PD** in specific articulation segments. Very similar features and methodology are employed in [124] to predict **UPDRS**, obtaining Spearman's correlations of 0.72 with the scores of the patients, using three databases: one in Spanish (50 patients), other in German (88 patients) and other in Czech (20 patients). The novelty of this last work consists in the inclusion of new additional features, the speech intelligibility measures, calculated in terms of word accuracy obtained with the Google speech recognizer.

Other works referred previously in this section, as [109] uses the velocity and acceleration of the amplitude envelope of the speech signal during a **DDK** test to obtain several complexity measurements to discriminate patients from controls (50 subjects per group). In the same manner, these features are used in a **SVM** classification scheme providing an accuracy of 85%.

Among all the articulatory approaches, some of them use vowel segments during articulation to obtain phonatory features related to the formants obtaining much better and more reliable results than when using sustained vowels, as it was described in section 4.3.3 [34], [55],

[95], [96], [114], [115]. Although all of them could be used to automatically detect or assess PD, only some include machine learning techniques or classifiers in the study.

Regarding severity prediction, the work [125] uses Chi-Squared Distance (CSD), proposing that this feature can provide a better correlation (0.78) with the UPDRS speech symptom severity ratings (UPDRS-S) than other traditional measurements such as HNR or MFCC, although an analysis of the correlation of all the used features with other UPDRS ratings such as the motor part would have been of interest. The work [126] combines a set of acoustic features obtained employing OpenSMILE and other features extracted from the analysis of articulators displacement using several read sentences, words and DDK tasks from the GITA database (50 patients) [127]. The novelty of this work is that the authors obtain the articulators displacement by inversely mapping the speech signal using Deep Neural Network (DNN). The Pearson's correlation between UPDRS and estimated scores (using Support Vector Regression (SVR) and DNN) reaches values of 0.51. Other examples of UPDRS estimation using similar acoustic features can be found in [128]–[131].

Lastly, and concerning automatic assessment, a recent study, [132], uses the phonetic posteriors obtained through the use of a DNN scheme to characterize parkinsonian voices as a blend of non-modal voices, suggesting that PD can be represented by a voice quality spectrum composed of 30% breathy voice, 23% creaky voice, 20% tense voice, 15% falsetto voice and 12% harsh voice.

4.4 PROSODIC ASPECTS

4.4.1 *Evidences of the influence of PD in prosody. The main cues: fundamental frequency and speech SPL*

Attending to the works found in literature about the prosody of parkinsonian speakers, two main features, fundamental frequency and speech SPL and their related features are analyzed in a great amount of works, since these are the two of the most characteristic acoustic traits associated to that aspect.

Although the early studies alluding to prosody on parkinsonian speech describe it as monotonous and with reduced loudness [133], one of the first articles performing objective measurements to characterize dysprosody or aprosody on patients, [134], does not show relevant differences in speech SPL (mean values and variability) between the 17 patients and 17 studied controls but a higher fundamental frequency and a reduced pitch range in patients. In a subsequent work, [135], authors perform acoustic measurements of prosody in a read sentence and spontaneous speech from 10 male patients and 10 male controls showing a higher fundamental frequency and a reduced speech SPL range in patients.

An early attempt to perform prosodic measurements in untreated patients can be found in [87]. In this work, authors measure frequency and speech SPL variability on sentences of 22 PD patients and 28 controls, finding a phonational range and dynamic range at natural frequency similar for both, patients and controls. However, this work only uses a read sentence for the prosodic analysis. Equally, in a questionnaire administered to each patient, they claim to suffer from monopitch, low SPL and tremulousness. The results on [87] are in concordance with those obtained by authors of [136], who do not find significant differences between groups (7 patients, 12 controls) using declarative, affectively neutral sentences.

To this regard, authors in [137] find elevated mean and reduced standard deviation of fundamental frequency in male speakers while no changes in mean and a significantly reduced standard deviation of fundamental frequency is found in female speakers. Here, four complex sentences are employed as speech task including a main clause and a subordinate clause. Fundamental frequency standard deviation is used in [58], obtaining similar conclusions.

In contrast, other works report more clear differences such as [6], in which authors employ diverse intonations of vowels (rising and falling intonation) from 20 patients and 15 controls, reporting a more restricted range of fundamental frequency in patients in comparison to controls. This study employs several linguistic modes in the used sentences as neutral, declarative, interrogative and imperative, in order to include distinct types of prosodic ranges. In fact, the differences between groups were only observable in declarative and interrogative modes. In this sense, the work [138] also finds a reduced fundamental frequency standard deviation for patients (10 subjects) respecting controls (20 subjects) specially in interrogative sentences. A reduced fundamental frequency and speech SPL variability is found too in [88], specially in subjects with advanced disease, using in this case monologues as speech tasks. In this same work, the maximum phonation frequency range extracted from a singing scale task is the feature that best differentiates between patients (in both, early and advanced stages) and controls.

In the same sense, [139] reports reduced speech SPL ranges in a group of 16 patients respect to 16 controls. In this case, speakers produce utterances in several emotional modes (neutral, happy and sad).

It is convenient to notice that these prosody indicators, as many others in literature have not been studied in the prodromal period⁸. To this sense, [140] makes use of the recordings of two individuals (one with PD and a matched control) appearing on an American television news services during 11 years. The analysis of the recordings, including 7 years of pre-diagnosis samples and 3 years of post-diagnosis, reveals

⁸ The prodromal period is the time in which the early signs start to appear but almost no symptom is appreciated by the patient or the clinician and the disease has not been diagnosed yet. It can be considered as the *incubation* period.

that changes on fundamental frequency variability can be noticed 5 years prior to diagnosis.

Regarding the combination of speech SPL and fundamental frequency ranges, [141] suggests the use of phonetograms to compare patients and controls. Figure 4.1 depicts a phonetogram of a patient and a control obtained after asking patients to perform different tasks as phonation of notes and glissandos among others. Although the phonetograms can be obtained using other speech tasks as fixed texts and monologues (i. e., producing different patterns), these figures are included here as an example of the potential of this tool, which can be used in automatic detection systems and, at the same time, can provide visual and easily interpretable information to clinicians.

(a) Control speaker

(b) PD patient

Figure 4.1: Phonetograms of a control (a) and a parkinsonian (b) speakers obtained from [141]. Gravity centers, marked with a black bullet, are located at $\{220\text{Hz}, 56\text{dB}\}$ for the control and at $\{196\text{Hz}, 57\text{dB}\}$ for the patient

Other features related to fundamental frequency, as the RFF, are found to be significantly reduced in patients (in ON and OFF states) with respect to controls [142], feature which tends to be lower as the disease evolves. This measurement reflects the variability of the instantaneous fundamental frequency around voiceless consonants (calculated using 10 cycles) respect to the reference fundamental frequency (calculated with a higher number of cycles, typically, 200). Therefore, a reduction of RFF implies a reduction of prosodic variability around articulation instants. In the same manner, other associated sign found in PD patients is related to difficulties to appropriately control the speech loudness in presence of background noise [143], in contrast to controls,

who can produce a higher speech SPL in accordance to the background noise (mechanism known as Lombard effect [144].)

With these premises, it is possible to claim that a considerable number of works find disprosodic characteristics in parkinsonian speakers related to speech SPL (mainly, reduced overall level and reduced ranges and/or variability) [35], [50], [88], [112], [113], [134], [135], [139], [143], [145] and to fundamental frequency (mainly, reduced ranges and/or variability) [6], [58], [88], [110], [134], [137], [140], [142], [146], [147]. However, some other works such as [87], [136], [148] do not find any fundamental frequency variability difference. Nevertheless, this variability seems to be highly dependent on the speech task used, and this is a key point. From the works not finding differences between groups to this respect, [136] employs 12 declarative and affectively neutral utterances, [87], a single neutral sentence and [148] only uses the repetition of a short neutral sentence while the rest of the cited works, which found significant differences in fundamental frequency variability between parkinsonian and controls speakers, include monologues or sentences in different linguistic modes (e.g., imperative or interrogative) or emotional modes (e.g., happy, sad or angry).

4.4.2 *Speech rate in parkinsonian speakers*

An early comparison of the speech rate between parkinsonian and control subjects is included in [134] where the mean speech rate found in both groups is similar, being the standard deviation higher in the parkinsonian group. This is caused by the existence of some patients with a very low and others with a high speech rate compared to the control group, which is more homogeneous in this regard. Subtle differences were found in the speech rates between groups in [149] where it is observed a tendency to a lower mean rate in patients in comparison to controls. The work [118], where the author analyzes the speech rate in conversational speech of 12 patients, and 10 controls, suggests that the differences on speech rate of both groups are not significant. Accordingly, Sabine Skodda et al. report no significant differences in overall speech rate between 121 patients and 70 controls [150], findings which are supported by other works such as [35]. Authors consider that these conclusions differ from other previous studies since the comparison of results in literature is inconsistent due to the use of small databases or different methodologies. However, in the same work, a higher acceleration at the end of long sentences in patients and a significant reduction of the total number of pauses are reported. Nevertheless, this same study points out a problematic that is left unattended in most of the works: no depression analysis was made in any of the patients and this could influence the results.

More recently, [151] proposes a new algorithm to measure speech rate in dysarthric speakers. By using this algorithm, authors observe an increased pace acceleration in PD patients respect to controls.

In general, it can be said that there is a certain controversy in literature with reference to speech rate, as different works [35], [118], [134], [146], [147], [149], [150], [152], report higher, lower or the same rate in parkinsonian with respect to control subjects.

Finally, as it is concluded in the review published by Blanchet et al. about the deficits on speech rate of parkinsonian speakers [33], their speech rate is different in most of the cases. It can be faster, or slower respect to controls and, frequently, with a higher variability. Additionally, authors identify in patients difficulties to modify the speech rate when necessary. In a posterior review [153], Skodda supports the same idea and suggests that the use of speech rate can be useful to monitor the evolution of the disease, as this feature shows a deterioration during the course of the disease.

4.4.3 *Other prosodic indicators*

Other studies focus on characteristics directly linked with prosody such as breath control, emphasis of syllables during speaking, pause duration, or even perception of prosody.

Regarding breath control, in [154] breath pauses of parkinsonian speakers are found in unusual moments of speech (agrammatical boundaries) as well as speech initiation at low lung volume in some patients, in comparison to the control group. In this sense, the work [148] also reports unorthodox pauses. This same work finds longer pauses between each sentence in a sentence repetition task and a lower percentage of speech time for each repetition in a study with 20 patients and 19 controls.

In work [45], perceptual and acoustic differences using the CAPE-V protocol [77] are found between 19 patients and 18 controls where it is possible to emphasize the observation of abnormal somatosensory laryngeal function in patients, associated with the timing of phonatory onset, respiratory driving pressure or lung volume employed per syllable. The study concludes that PD produces somatosensory laryngeal deficits which can produce phonatory, articulatory and prosodic abnormalities, as these somatosensory mechanisms -as well as the auditory- provide the feedback to produce the accurate adjustments necessary for a correct phonation and articulation. In other words, there are some deficits in the integration in the basal ganglia of the stimuli received in the sensory cells during speech -or swallowing- deriving in a malfunction of voice and speech mechanisms. Subsequently, the study [45] concludes that the abnormalities in the laryngeal somatosensory function produce deficits in the control of pitch and voice intensity. The study points out that these abnormalities may also increase the sensitivity of the auditory system during the speech, magnifying the self-perception of the vocal intensity and resulting in a decreased sound pressure level.

Respecting to accentuation and phonemic stress⁹, in the work [155], 30 patients and 15 controls are asked to perform prosody comprehension tasks and to read several sentences as well as to describe an image. No differences in prosodic comprehension were found between the two groups. However, patients demonstrated to have problems with the mechanisms involved in the pronunciation of noun compounds. In general, controls make a clear distinction between noun phrases (e. g., green house) and noun compounds (e. g., Greenhouse) while patients failed to perform the prosodic actions (authors measure, mainly, pitch and pauses between words) to clearly differentiate the two types of utterances. Similar results in the prosodic perception by 11 patients are obtained in [156] where there are no significant differences between the two groups in phonemic stress perception. However, this work demonstrates a general reduction in the ability of patients to identify the linguistic and affective meaning of sentences in comparison to controls, coinciding with [157] and [158]. To this respect, this last study concludes that *"the basal ganglia provide a critical mechanism for reinforcing the behavioral significance of prosodic patterns and other temporal representations derived from cue sequences facilitating cortical elaboration of these events."* In agreement with these conclusions, [159] presents a higher percentage of mismatch in the parkinsonian group (14 subjects) than in controls (14 subjects) in the recognition of emotions.

4.5 LINGUISTIC ASPECTS

The linguistic aspects, unlike the rest of the aspects already mentioned, belong to the so called non-motor symptoms. In this case, acoustic measurements on speech are mainly reduced to pauses length. To automatically analyze the linguistic aspects on patients using speech, the use of Natural Language Processing (NLP) is required in order to find repetition of words, grammatical structure or vocabulary.

In general, literature suggest that the linguistic distortions are mainly present in the advanced stages of PD [160]. However, works like [161], examining the self and carers perception on communication skills reveals a negative influence of PD on these skills, independently of the age and of the stage of the disease. Attending to the survey scrutiny, this influence is perceived even before the motor symptoms and other speech impairments appear. These findings are supported by a recent study [162] in which authors see a possible existence of a sui generis cognitive deficit since the early stages of PD.

⁹ The phonemic stress can be defined as the relative emphasis given to one syllable respect to the others within a word, which can be caused by changes in the fundamental frequency and an increased loudness as well as a more clear articulation and increased vowel length. In some contexts, the term stress is substituted by accent.

4.5.1 *Disfluency*

One of the first representative works addressing the influence of PD on the linguistic aspects of the speech production is [135] where patients exhibit a higher number of silent hesitations per time unit, longer silent hesitations, more words per silent hesitation, and a decrease in the number of modalizations and interjections. Equally, results suggest a categorization of patients into distinct levels of severity based on the number of hesitations per minute and in the syntactic complexity. As a conclusion, authors consider that *"changes in the structure of spontaneous language production with increasing severity of dysarthria reflect PD patients' adaptation to their disease"*. In a subsequent work [163], the same author finds language structure disruption in patients respect to controls, consisting in long-duration silent hesitations. The same work reports that patients tend to develop adaptive strategies to counteract the verbal difficulties, being less effective as the disease advances.

The work [160] studies 10 patients and 10 matched controls, who are asked to answer several questions during an interview. The speech is recorded and analyzed to obtain the proportion of grammatical utterances and complex sentences respect to the whole discourse, among other features. The findings indicate that the productive syntax of patients is correlated to dysarthria.

Attending to word repetition in patients, the publication [164] finds hyperfluent (palilalia) or disfluent (stuttering) variants of word repetition in 28% of the patients (53 patients studied), most of them in advanced stages of the disease. A more recent work, [165] describes the same stuttering-like disfluencies in a study of 32 patients where the level of disfluency is found to be higher during the production of monologues respect to reading tasks. The work [166] finds significant differences between patients and controls too in different speech tasks, being more common in conversation than in other task conditions such as reading, repetition, singing or recited speech.

The publication [167], studying linguistic aspects of the speech of 16 parkinsonian (mild to moderate) and 16 matched control speakers, reports a longer time to initiate speech in patients as well as more errors in the proposed speaking tasks in comparison to controls. In the same manner, patients show a decreased capability in certain receptive language tasks.

4.5.2 *Use of action-verbs*

In general, literature demonstrates that the linguistic problems on patients affect to speech production and perception. In general, the works about comprehension of language of PD patients are more abundant as the scoring or evaluations of results are easier to perform and more reliable. For instance, the work [168] proposes a methodology to detect PD by measuring language comprehension abilities of patients. In the

study, in which 22 patients and 22 controls are enrolled, authors ask the subjects to pair several sentences, with different levels of complexity, with certain images, containing actions and objects. The scoring of the subjects is performed as a function of the number of errors in the pairing and then these values are introduced into a logistic regression model, obtaining an accuracy of 84% in distinguishing patients from controls. Another example of this type of analysis is [162] where authors find action-naming and action-semantic deficits in patients, meaning that the main impairments are associated to the use (comprehension or during speaking) of action verbs.

This is confirmed in works like [169], [170] where patients have a significant impairment and a longer time response in comparison to controls when naming actions in pictures. These impairments are less pronounced when naming objects.

4.5.3 *Other approaches*

There exists a great variety of works analyzing the linguistic aspects of parkinsonian patients from different points of view, finding in some cases, certain particular behaviours associated to the disease. For instance, the work [171] finds similarities in the clinical characteristics between parkinsonian patients and aphasics respect to the phenomena called "tip-of-the-tongue" and "word production anomia", related to the word-finding difficulty but unrelated to memory loss or intellectual impairment.

The work [6], analyzes prosodic, articulatory and cognitive aspects in several speech tasks and signed language tasks of parkinsonian speakers and signers, and points out that patients are highly sensitive to cognitive load, and this load finally influences not only articulation or prosody but other linguistic aspects related to the content of the message. In this sense, the work [172] finds a lower proportion of formulaic expressions¹⁰ in patients (16) respect to controls (18). These results are confirmed in [166].

A different approach, [173], performs microlinguistic and macrostructural measures of the speech of patients during several tasks such as story retelling and image sequence description (19 patients, 19 controls). Using discriminant function analysis, authors obtain an 84% of accuracy on differentiating patients from controls, revealing significant differences between the two groups in grammar, informativeness and verbal disruption.

¹⁰ Formulaic expressions are speech formulas or fixed phrases, usually employed to express greetings, agreement or apologies among others. Examples of formulaic expressions are: "excuse me" or "thanks a lot"

4.6 COMBINED APPROACHES

Some of the most recent publications found in literature, utilize the knowledge generated along the years in the four mentioned aspects (phonatory, articulatory, prosodic and linguistic) and combine them to obtain, systems to automatically detect or assess PD through speech. An overview of those considered as the most relevant is included in this section.

In the work [40], authors say: *"It is important to note that the PD disorders of the individual subsystems not only influence each other but also frequently overlap. Disturbances of respiration and phonation consequently reflect, in particular, disruptions in speech prosody and partially articulation."* This work is an example of combination of phonatory, articulatory and prosodic aspects into an automatic classification system providing an accuracy on distinguishing PD from controls close to 90%. On their system, 4 phonatory, 5 articulatory and 10 prosodic features are selected from a larger group by studying the statistical significance and correlation between features, where statistically insignificant features are removed and only one in a group of measurements is maintained. Each of the selected features is used in a separate classifier based on the Wald task [174], where the output scores are -1 if the measurements match the control group performance, 1 in the case of PD and 0 for indecisive situations. Combining the results of each classifier, one per type of measurement, the global result will be the sum of all of the scores. Authors consider that speakers with global scores over 1 are considered to have any type of speech impairment, presumably attributable to PD. The favorable points of this methodology to be used in a clinical environment is that the measurements are easily interpretable and it is possible to have individual scores for each speech task, for each speech aspect and a global value. Additionally, the used database is highly relevant as it mainly contains speech from untreated patients in an early stage. The main drawback of the work is that there is not a strong validation methodology as the same data is used to select the features and train the models and, therefore, the models can be over-fitted to the database, so results could be optimistic.

A very similar methodology is detailed in the work [58] from the same authors, where 6 phonetic, 4 articulatory and 3 prosodic features are measured in different speech tasks on 19 patients before and after starting pharmacological treatment and in 19 controls. Additionally, in this case, results suggests that almost 90% of patients improve in, at least, one of the studied aspects of speech after treatment. Furthermore, an analysis of the correlation between these measures and UPDRS global scores reveals no statistically significant correlation. However, a significant correlation was found between bradykinesia subscores and vowel articulation features and between rigidity subscores and vowel articulation and intonation features, in all cases in a monologue speech task.

A combination of speech tasks and articulatory, prosodic and phonatory features is employed in [175] along with an automatic classification system for PD detection. In this work authors use several speech tasks from 88 patients and 88 controls, all of them German speakers. The articulatory features consist in Gaussian Mixture Model (GMM) mean vectors (one per utterance) obtained from MFCC. Prosodic features are calculated from voiced segments and include statistics (mean, maximum, minimum and standard deviation) of a group of typical acoustic measures including fundamental frequency, jitter, shimmer, relative speech SPL and duration of the segment. Phonatory features consist in the parameters of a physical glottis approximation of a two-mass vocal folds model described in [176]. Additionally, 1582 features from OpenSMILE are used too as a brute-force approach. All of these features are included in four separate classifiers using SVM with a linear kernel. Each type of features is studied separately and an unweighted score-level fusion is performed to evaluate the fusion of the four systems (three speech aspects). Using a LOO validation scheme, results of 82% of accuracy are obtained using the articulatory approach and no improvements are reported by the fusion of systems. By using the same features and speech tasks in this work, authors try to predict UPDRS scores divided into only three levels (low, medium and high) obtaining accuracies no higher than 60%.

On [123], authors automatically segment the speech signal into voiced and unvoiced, performing prosodic, phonatory and articulatory measurements over running speech which are introduced into a SVM classification system to separate parkinsonian speakers from controls in different languages. Three different systems are tested, one for each language (Spanish, German and Czech) obtaining accuracy results ranging from 60 to 99%. On spite of the good results in some of the languages, it is considered that these are over-optimistic *"due to an over-fitting of the model, since it was optimized during training"* in subsequent words of the same authors [177].

A more recent contribution is the proposal of an open-source software for parkinsonian speech analysis, namely "Neurospeech" [178]. This software allows to calculate and analyze several phonatory, articulatory and prosodic features as well as assessing intelligibility and predicting UPDRS-III score and dysarthria level of speech according to the Frenchay Dysarthria Assessment (FDA-2) score [179]. These features are state-of-the-art characteristics such as shimmer, jitter, fundamental frequency statistics, τ VSA or FCR among others, and it is possible to compare the calculations from a new voice or speech recording with other values obtained from healthy controls. In this first version, these control values are obtained from the subjects of the GITA database [127] but as it is an open-source code, developers can modify these reference values with others originating from a larger database or after a clinical validation. Thus, instead of proposing new methods or techniques in the detection or assessment of PD from speech, or includ-

ing a combination of different aspects with these purposes, this work proposes a transference of the research of multiple studies into a tool, allowing clinicians to easily evaluate several features at the same time and to visually compare these results with a reference.

Finally, most of the works performing severity prediction (UPDRS or H-Y), where [42], [180] are two examples, use a blend of features and/or tasks which can be considered a combined approach. For instance, [42] obtains a mean absolute error of 5.5 on the UPDRS of 168 patients, suggesting that the reading task is more appropriate for the prediction than sustained vowels or DDK tasks, using MFCC, jitter, shimmer, and other features calculated with openSMILE.

4.7 ANALYSIS OF DATABASES FOUND IN LITERATURE

4.7.0.1 *List of relevant databases*

In this section, a review of several parkinsonian voice and speech databases is performed. Table 4.1 includes a list of the most relevant databases mentioned in this thesis, including the first work referring to it, the year of publication, the language of the database, the number of patients and controls, the type of parkinsonism of the patients and the speech tasks recorded. One of the reasons why this table is included here is to provide the reader with an extended information about the cited databases, making it easier to establish a comparative of the characteristics of each one and, hence, of the works using them and the conclusions extracted from these works. The information included in the table is referred to the database used in the specified work. However, some of the works refer to these same recordings in posterior works and cite the original publication to describe the database but then use a different number of patients and/or speech tasks, not present in the original document. For instance, the work [137] describes for the first time a database containing 138 patients and 50 controls from which authors recorded four read sentences. However, the subsequent work [123] uses this database and refers to it as "*the database described in [137]*" but in this new study 88 patients and 88 controls are part of the corpus while performing more speaking tasks as sustained vowels or monologues. This means that, although Table 4.1 makes a description of some databases, in some cases authors expand their characteristics in posterior works.

It is important to notice that, to the knowledge of the author of this document, almost no database of the listed is publicly available. Only [91] can be downloaded from a website. However, it is possible to find some authors willing to collaborate and to share the recordings that they retrieved under certain conditions or in exchange of a certain compensation.

Year	Work	Language	Number of speakers	Pathologies	Type of material
1978	[48]	American English	200 patients	idiopathic and postencephalitic PD	Read sentences and 3 to 5 mins of conversational speech
1988	[135]	American English	10 patients, 10 controls	PD-not specified	Read passage and spontaneous speech.
1988	[155]	American English	30 patients, 15 controls	idiopathic	16 noun Compounds and 16 paired Noun Phrases. Image description.
1993	[136]	German	12 patients, 12 controls	idiopathic (7), Huntington's Disease (HD) (5)	12 declarative and affectively neutral utterances.
1997	[145]	American English	30 patients, 14 controls	idiopathic	Sustained vowels, monologue and picture description.
1997	[87]	Spanish	22 patients, 28 controls	not specified	A sustained vowel and a read sentence
1998	[138]	French	20 patients, 20 controls	idiopathic (10), Friedrich's Ataxia (10)	40 read sentences
1999	[50]	Australian English	200 patients	idiopathic	Two minute conversational speech
2000	[88]	Australian English	60 patients, 30 controls	idiopathic	A sustained vowel, a singing up and down scale, 1 minute monologue.
2005	[154]	American English	7 patients, 6 controls	idiopathic	Monologue. Te recordings include additional data: measurements of kinematic information of the chest wall to determine respiratory movements
2008	[150]	German	121 patients, 70 controls	idiopathic	Four complex sentences
2008	[139]	German	16 patients and 16 controls		Repetition of a word in different emotional intonations
2009	[100]	American English	52 patients	not specified	Sustained vowels and two picture description: with and without distractions. Although in the cited work, authors only refers to these patients and tasks, posterior studies citing this database (such as [42])include much more patients and tasks.
2011	[137]	German	138 patients, 50 controls	idiopathic	four sentences
2011	[167]	American English	16 patients and 16 controls	idiopathic	Six sentences varied in both, length and syntactic complexity.
2011	[67]	American English	10 patients (with DBS) and 12 controls	idiopathic	DDK.
2012	[66]	American English	32 patients and 32 controls	not specified	Rainbow passage and three repetitions of a read sentence. Patients are recorded in two sessions. One at least 9h after medication intake (OFF) and other 45 min after medication intake (ON)
2013	[34]	German	80 patients, 60 controls	idiopathic	A sustained vowel and four sentences
2013	[117]	American English	15 patients, 15 controls	PD-not specified (8), Multi-system atrophy (1), Corticobasal degeneration (1), Friedrich's ataxia (3), Cerebellar toxicity (1) and unknown (1).	Syllable sequences, ranging from low to high complexity.
2013	[95]	Czech	20 patients, 15 controls (only male)	idiopathic	Sustained vowels (/a/, /i/, /u/), sentence repetition, read passage and monologue.
2013	[58]	Czech	19 patients, 19 controls	idiopathic	A sustained vowel, four sentences, a DDK task (/pa-ta-ka/) and a monologue.
2013	[91]	Turkish	20 patients, 20 controls	not specified	26 samples, including sustained vowels, words, and short sentences.
2013	[112]	american English	24 patients, 15 controls	13 idiopathic, 11 multiple sclerosis	25 read sentences in habitual, clear, loud and slow conditions.
2014	[127]	Colombian Spanish	50 patients, 50 controls	idiopathic	Sustained vowels, DDK tests, isolated words, read sentences, read passages and monologue.
2015	[148]	Italian	20 patients, 19 controls	idiopathic	10 repetitions of a specific read sentence.
2017	[181]	Czech	48 patients, 16 controls	Idiopathic PD, progressive supranuclear palsy and multiple system atrophy	Several series of syllable sequences.

Table 4.1: Parkinsonian speech databases identified in literature.

4.7.0.2 *Recording conditions*

The recording of a voice and speech database to study the influence of a pathology in the voice of the subjects must obey some rules to avoid an excessive influence of the channel in the audios, leading to biased studies or detectors in some cases. For example, some databases such as [95], [99], [127] have been recorded in sound-treated booths. The fact that the booth is acoustically isolated does not mean that the acoustic conditioning, and therefore, the acoustic field inside the booth is the most adequate to record voice signals. Mainly, these cabins are designed to perform audiometries, specially if its volume does not exceed 8m^2 and include a window. Therefore, it is important to consider all the acoustic elements which could affect to the recording, from the microphone to the acoustic environment.

A major drawback of the databases used in some of the revised publications like [97], [99], is that recordings are performed using portable and distinct equipment introducing noise and variability in which could bias the system. However, the advantage is that these databases usually contain a large amount of utterances and are expected to be more suitable for approaches employing machine learning technologies.

In the same way, some of the works such as [87] include a previous analysis of the larynx of the speakers performed by a specialist or a hearing test, like in [114], in order to discard any organic pathology or hearing impairments which could influence the phonation or the speech. However, most of the works only incorporate a survey to discard smokers or subjects with known vocal issues. Others such as [155], perform a study of depression which is commonly associated to PD and, in the same way, affects prosody.

After examining a large amount of works in the literature, it is possible to observe that many of them use speech corpora from patients under pharmacological treatment, and one logic question could be if these treatments are influencing the results. Depending on the study, the answer can be affirmative but with a high uncertainty. To this respect, one of the more representative and recent studies is [58] in which authors objectively analyze phonatory, articulatory and prosodic features of 19 patients before starting a symptomatic treatment and between one and two years after the beginning of the dopaminergic therapy. A statistical analysis of these features reveals no significant differences between the two recording sessions. However, a second analysis employing pattern recognition techniques shows that some patients tend to be classified as controls specially when utilising some specific features related to speech SPL, voice quality, fundamental frequency variability and articulation. Other works show no influence of dopaminergic therapy on articulatory features [142].

4.7.0.3 *Subjects and materials*

The number of speakers included in the studied databases ranges from 2 to 200. It is difficult to extract a clear conclusion on which is the minimum recommended number of subjects that a database should include since it all depends on the goals of the study using the recordings and the speech tasks to be used. What is clear is that the articles using a low number of subjects, such as [140] serve as case studies but their conclusions cannot be considered as generalist until other works using more subjects can confirm them. In contrast, some collections of recordings such the one used in [40], while containing a relatively small number of speakers (46, from which 23 are patients), are highly relevant as these include only newly diagnosed patients. In this database, none of the patients is under treatment, and the mean time since diagnosis is 30 months.

The most important differences among different studies analyzing the same aspect are, usually, the number of subjects, the stage of the disease and the type of speech tasks. These differences are the main source of discrepancies in the studies, as it is stated in [150]. For instance, works containing interrogative sentences or in any kind of emotional mode, usually find more influence of PD in prosody than studies employing neutral sentences as suggested in some works [6], [139].

Considering the use of a database for the detection of PD using speech processing and machine learning technologies, the duration of the recordings should be long enough to be representative of the speech or voice of a person. For instance, typical lengths to recognize the speaker with a high accuracy in speaker recognition technologies are around 1 min. In the same manner, most of the works present the speakers groups (usually, parkinsonian and control) matched in number, age and sex. This is crucial since age and sex influence the acoustic information of voice and speech by itself and, to this respect, significant differences between patients and controls in a corpus could bias the results.

4.8 LIST OF FEATURES FOUND IN LITERATURE

Table 4.2 includes a list of some of the motor feature families and coefficients analyzed in the literature to detect or assess PD, containing the behavior of these features in parkinsonian voices reported in literature and some of the works that use them. All of these features have been already presented in this chapter and the most relevant are discussed in the following section.

4.9 DISCUSSION

After a review of the literature, four main speech aspects influenced by PD are identified: phonatory, articulatory, prosodic and linguistic.

Aspect	Family of features	Influence of PD	Works including the family of features
Phonatory	Jitter	Usually, reduced values of jitter are found on patients respect to controls.	[85], [87]–[89]
	Shimmer	Usually, reduced values of shimmer are found on patients respect to controls.	[40], [84], [85], [87]–[89], [34], [58], [90], [91], [94], [97], [42], [98], [99], [175], [178]
	Maximum Phonation time	Results suggest that this value is lower in patients respect to controls.	[84]
	HNR	Results suggest that this value is lower in patients respect to controls.	[40], [58], [85], [87], [91], [97], [98], [99], [125]
	NHR	Results suggest that this value is higher in patients respect to controls.	[34], [88]–[91], [94], [97]–[99]
	Subglottal pressure	Results suggest that this value is higher in patients respect to controls while producing the same speech SPL	[45], [83]
	PPE	Results suggest that this value is higher in patients respect to controls.	[58], [97]–[99]
	RPDE	Results suggest that this value is slightly higher in patients respect to controls.	[58], [97]–[99]
	MPT	Results suggest that this value is slightly lower in patients respect to controls.	[58], [97]–[99]
	D_2	Results suggest that this value is higher in patients respect to controls.	[89], [97]–[99]
	SPI	Results suggest that this value is higher in patients respect to controls.	[90]
	Tremor	Results suggest that this value is higher in patients respect to controls.	[48], [82], [87], [90], [97], [182], [183]
	CSD	Results suggest that CSD obtained from sustained vowels can provide a high correlation with UPDRS-speech	[125]
Articulatory	VAI	Significant reduced values of VAI are found on patients respect to controls.	[34], [93], [95], [114]
	VSA	Reduced values of VSA are found on patients respect to controls.	[55], [95], [112], [114], [115], [93], [94], [178], [184]
	VOT	Result suggest that VOT tends to be longer in patients respect to controls.	[40], [110], [181]
	MFCC	These features are usually used for automatic detection and assessment of PD.	[42], [99], [104], [123], [125]
	PLP	These features are usually used in some spare works for automatic detection and assessment of PD.	[102]
	Duration of voiced segments	This duration tends to be longer in patients.	[6]
	Voice bar length	This duration tends to be shorter in patients.	[6]

Table 4.2: Some of the speech motor feature families found in literature with its behavior and works including them.

After this review, one of the main observations is that, although the number of works of each group is abundant, there are not too many approaches using automatic classifiers to detect or assess the disease or, at least, providing a score relative to the likelihood of the speaker to suffer from PD. However, many of these works provide evidences and suggest trends about the distinct features that can provide biomarkers or can be used in new tools for the diagnosis of PD.

4.9.1 *Phonatory approaches*

After a review of the phonatory approaches, it is observed that amplitude and frequency perturbations (jitter and shimmer), as well as noise, and their associated features have been widely used along time. Although literature demonstrate in some cases the influence of PD in jitter and shimmer [34], [40], [58], [85], [87], [90], [94], some works suggest that these features do not serve as indicators of PD or do not provide a significant differentiation between controls and patients [84], [88], [89].

In the case of noise, most of the works analyzing its presence on the parkinsonian voices indicate a relationship between noise features and the presence or evolution of PD [34], [40], [58], [85], [87], [90], [94], [97], [99].

It is difficult to extract a precise conclusion respect to some acoustic measurements since different works include a different number of patients in distinct stages of the disease and with dissimilar treatments. But in general terms, a considerable number of works point out the presence of incomplete closure of the vocal folds in patients, producing a breathy voice and, thus, a voice with more presence of noise (lower HNR) whereas jitter and shimmer seem to be significant more sporadically.

Nevertheless, in general, caution must be taken when using some acoustic features to detect or assess PD since these are not only influenced by the disease but by other factors as age [185], [186]. Therefore, these values must be assessed according to the age or the sex of the subject since the ranges of normality may vary. In the same manner, although it could result obvious, the existence of any other organic pathology influencing these features must be discarded in advance.

Other works use complexity measurements or a combination of a large number of features including cepstral and frequency coefficients or noise in automatic detection systems with unequal results.

4.9.2 *Articulatory approaches*

In general, it can be stated that most of the publications point out that the movements of the articulators and the structures involved in articulation (lips, tongue, jaw or soft palate) are reduced respect to

normal articulation or fail to fully contact during constriction. At the same time, the timing of the vocal folds closure is imprecise in some occasions, in combination with the rest of articulators during speech. Therefore, literature suggests that the combination of these deficiencies can cause abnormal voiced and unvoiced segment boundaries.

The articulatory impairments are mainly concentrated in obstruent consonants, plosives and fricatives (specially velars), the consonants requiring the greatest narrowing or closure during articulation. Another characteristic found in several studies is the appearance of spirantization in patients, a phenomenon consisting in the substitution of plosives with fricatives, product of the misarticulation of these last consonants. However, these phenomena are not clearly documented and measured objectively.

Several works studying the velocity and acceleration of the articulators suggest that some patterns obtained from these kinematic features can reveal the presence of PD [35], [61], [109], [118]–[120]. Provided that one of the most characteristic signs of patients with idiopathic PD is related to impairments in acceleration and velocity of movements (as akinesia, bradikinesia and hipokinesia), these findings are consistent. However, the analysis of velocity and acceleration of articulation based on the speech signal has not been deeply analyzed.

On the other hand, a considerable number of works study the VSA and other related features obtained from the vowels of running speech from patients and controls, reporting in most of the cases a reduced area for patients, indicating that there exist a reduction of the tongue elongation during the articulation of vowels.

In general, works performing automatic detection of PD taking advantage of these observations combining articulatory features and machine learning techniques, typically yield accuracies over 80%.

4.9.3 Prosodic approaches

One of the main findings related to parkinsonian prosody is that a reduction of speech SPL or a reduction of its variability are reported in multiple studies as the primary prosodic impairment associated to patients [35], [50], [88], [112], [113], [134], [135], [139], [143], [145].

Other feature family appearing in literature with remarkable results is the one related to fundamental frequency variability during speech, where most of the works report a reduced range in patients respect to controls [6], [58], [88], [110], [134], [137], [140], [142], [146], [147].

Regarding these frequency variability, measured from speech or phonatory tasks, there are almost no studies using singing tasks (the only work found in literature is [88]). This type of tasks could be useful since it requires changes of fundamental frequency, although intonation capabilities could be different in distinct individuals independently of their neurological state. To mitigate these issues, some speech therapists trying to evaluate fundamental frequency transitions and ranges

in patients, use well-known songs such as *happy birthday*¹¹ with the cited purposes.

With respect to the used materials, speech tasks containing different linguistic modes such as interrogative or imperative or emotional modes such as angry or happy provide, usually, the best differentiation between patients and controls in most of the analyzed features, specially in those related to fundamental frequency and speech SPL.

In the same sense, the use of noun compounds and noun phrases (Greenhouse vs. green house) can provide an extra information in the analysis of prosody of parkinsonian speakers.

Finally, literature suggests that the variability on speech rate in groups of patients is usually higher than in controls. However, there exist a certain controversy on the use of this type of features since patients have a higher or lower speech rate than controls but there is not a clear pattern to this respect.

4.9.4 *Linguistic approaches*

The studies in literature proposing linguistic approaches to differentiate parkinsonian patients from controls are focused, unlike the other works mentioned in this chapter, in non-motor symptoms. The linguistic distortions are manifested, principally, as disfluencies such as long silent hesitations, stuttering or palilalia and problems in production and perception of action verbs. In this case, there are not too many approaches using machine learning techniques and automatic detectors. According to the literature findings, these systems should carry out an analysis of the speech produced by patients after watching certain images. This analysis would involve the study of the type of vocabulary, response times, word repetition or hesitations among others and the resulting features could be included into supervised systems. In this sense, the possibility of using unsupervised approaches would involve the analysis of interjections, vocabulary, grammar complexity or formulaic expressions.

4.9.5 *Combined approaches*

Combined approaches bring together features from two or more different aspects into a single scheme, merging the discriminative properties of each aspect. The main hypothesis of these approaches is that these properties can be somehow independent and complementary and the combination can provide a synergistic scheme. Literature suggests that these types of approaches are the most appropriate to help clinicians on the differential diagnosis as these can cover different types of mo-

¹¹ The author of this text has observed this task in the protocols of speech therapists assessing patients with neurodegenerative diseases although almost no studies exploit this type of task.

tor (and non-motor, if linguistic aspects are considered) signs at the same time. To this respect, some of the latest systems proposed in the literature that combine several aspects allow to observe the results and features separately to permit clinicians to assess the influence of the disease on speech more precisely and to make more resonated decisions.

4.9.6 *Final considerations*

Although some early works [48], [149], [187] point out that PD has a clearer reflection on the phonation of patients than in articulation or prosody, more recent works [40] suggest that on early stages, prosody is the most affected speech aspect. Works like [140] support the idea that prosody is affected since the prodromal period, although authors did not study other aspects but prosody. At the same time, other publications like [161], [162] point out to cognitive impairments before the appearance of dysarthrophonic signs. Equally, studies analyzing articulatory aspects in newly diagnosed patients, provide good results in differentiating them from controls [123]. In general, although the influence of the disease in most of these aspects is noticeable in mid to advanced stages, there is no consensus on which aspect is more appropriate to help on differential diagnosis in early stages. Therefore, a potential solution is to use a combined approach, containing at least two aspects, for instance, articulatory and phonatory or articulatory and prosodic.

In the same sense, some studies point out that read sentences or monologues can help better to obtain an accurate differential diagnosis than other materials such as DDK tasks or sustained vowels [188]. However, these conclusions must not mislead research on avoiding the use of sustained vowels or DDK tasks, as the selection of the most appropriate material depends on the analyzed aspect and the feature to be measured. For instance, these two tasks can be performed, almost identically by patients with distinct mother tongues and, thus, can be employed in language-independent approaches.

From all the analyzed studies referred to each aspect, it is possible to identify some missing analysis which should be addressed. Although tremor is a common sign of PD, vocal tremor has not been deeply studied and used in phonatory approaches and only some works study it with heterogeneous results [48], [82], [87], [90], [97], [182], [183]. This could be caused by the difficulty to differentiate the parkinsonian tremor from other types of tremor such as microtremor or to the absence of reliable features related to this characteristic.

In the same sense, few works analyze the articulatory differences of some specific segments between parkinsonian and normal speech, such as plosives, fricatives or vowels. Moreover, these differences have not been widely exploited (and explored) in automatic detection systems. To this respect, some approaches using speech processing and

machine learning techniques aimed to detect and assess PD have appeared in the last decade, using the features and contributions provided by previous studies. Nevertheless, there are still several mature cutting-edge technologies related to speech and speaker recognition which have not been tested for PD detection or assessment.

It is important to remark that most of the works make no analysis of the differences of the influence of PD on voice and speech and the influences of other neurological diseases and this is important for diagnosis since in most of the cases, clinicians struggle to differentiate between two or three possible diseases. For instance, the work [120] shows a significantly reduced F_2 slope for PD patients respect to controls. However, the same results are obtained for stroke patients. Thus, some works analyze the differences between idiopathic PD and other neurological conditions such as ataxia, HD or multiple sclerosis among others [23], [31], [112], [113], [117], [136], [138], [149], [151], [163], [181], [188]–[191], but these are scarce and, up to date, insufficient to extract solid conclusions to be applied in the clinical practice. However, these types of studies are crucial, because, as it is remarked in [192], there exist common phenomena and differences on the speech production perturbations of the different motor speech disorders. Without a differentiation between the main neurological impairments on their influence over speech, it will be more difficult to propose an effective tool to detect and/or assess PD. For instance, [136] and [138] show clear different prosodic aspects between PD and HD and Friedreich's Ataxia patients, respectively and [181] reports articulatory differences in progressive supranuclear palsy, multiple system atrophy and PD, specially when analyzing VOT.

Regarding the different languages, not too many works address the creation of multilingual systems nor analyze the behavior of the proposed schemes in more than one language [123], [184], [193]. In general, different languages share some types of allophones and have some unique or language-dependant ones, so the articulatory analysis using certain tasks such as DDK could provide the same results for different languages. As it is stated in [193], if the approach is more focused on prosody the language specificities must be considered, as each language has its own prosodic structure considering lexical stress or pitch evolution.

Most of the systems created to distinguish between parkinsonian speakers and controls or to predict the patient's level of affection employing only a read sentence or a short passage can be considered as *supervised approaches* as the models obtained are dependant on the specific allophones contained in the utterances. Some of the approaches using specific segments from the speech could be considered *supervised approaches* too, as these features will depend not only on the segments used to obtain the features. Other types of systems can be treated as *unsupervised approaches* when employing speech recordings of long dura-

tion (usually, hours) with a high variability of lexicon, as monologues, in training and testing.

Concerning the accuracy of the reviewed systems performing automatic detection of PD, the most recent ones usually yield values over 80% and, in some publications, these values are over 95%. In these last cases, all the identified works were employing over-optimistic or unorthodox methodologies, diminishing the relevance of their results. Therefore, it can be stated that the most relevant and recent works provide an accuracy varying between 80% and 95%. In any case, it is important to question the possibilities to obtain accuracy values near 100%. Considering the analysis of works studying the number of patients with speech impairments introduced in section 2.7, it is clear that, at least 80% of them have signs of PD on their speech. However, the maximum portion of patients with these signs remains unclear. In the same manner, some studies point out that between 2 and 4% of patients diagnosed with idiopathic PD were from other type of pathology, as autopsies reveal [18], [194]. Moreover, this diagnosis is more precise as the disease progresses. This means that some of the patients included in a parkinsonian speech database could not be really affected by PD, being this more likely in the newly diagnosed patients. These two considerations would impose a minimum false rejection percentage in automatic detectors, to values which could be close to 10%.

Finally, although there are works studying the voice and speech of patients since more than 50 years ago, as it has been described in this chapter, it is not possible to find so far many clinical validation studies from the neurological service of any hospital using objective biomarkers or automatic detector systems employing speech as input source for the diagnosis of idiopathic PD (to the author of this thesis knowledge). These type of studies are aimed to validate the experimental biomarkers or systems in the true clinical practice and can entail a large number of patients and a long evaluation time. This fact may be attributable to the lack of tools providing clear indicators but not because speech cannot be a good candidate as source of information during diagnosis.

OVERALL METHODOLOGY AND MATERIALS

This chapter includes a description of the overall methodology employed in this thesis. Considering the four main speech aspects which are influenced by the presence of PD, namely, phonatory, articulatory, prosodic and linguistic, the first two are studied separately in this work, proposing new approaches and employing cutting edge speech technologies which have not been exploited, to the moment, for this task. Additionally, a combination of these two aspects is explored with the aim of using the potential complementary capabilities of the correspondent used methodologies for the automatic detection of the disease. In further chapters, each one of the experiments involved is explained in detail, enclosing not only the procedures followed in every case but a theoretical introduction to the techniques used.

Likewise, six speech databases were employed in this thesis: three main databases, including parkinsonian speech and three other auxiliary databases to create the Universal Background Model (UBM) [195] and Forced Alignment Model (FAM) [196] as detailed in their respective chapters. The description of these corpora is incorporated in this chapter.

5.1 ARTICULATORY APPROACHES

Speech and speaker recognition technologies are currently arriving to a mature state. Decades of work from thousands of scientists have brought to us multiple successful techniques which can extract very precise information from the speech in order to identify a speaker or to segment the speech to the phoneme level with high accuracy. Considering this, along all the articulatory experiments underlies the idea of taking advantage of these techniques, adapting them to the particular scenario of PD diagnosis support systems. In the proposed experiments, some of the most successful technologies related to speaker recognition (like PLP in the front-end or i-vectors for the classification phase) or speech forced alignment were analyzed in a different context: the detection of a neurological disease. Within each experiment, different approaches or variations of the same basis scheme are proposed in order to find the optimal solution.

In general, the identification of the proposed schemes as supervised or unsupervised and language-dependent or -independent systems relies on the material used in each case, as it is explained in the correspondent chapters.

5.1.1 *Experimental set 1: use of state-of-the-art speaker recognition technologies*

In this experiment, state-of-the-art speaker recognition techniques were applied and adapted to a new application domain: the automatic detection of **PD** using the patient's speech. Three families of features were considered, **MFCC**, Rasta-**PLP** and Linear Predictive Coding (**LPC**), along with their respective derivatives, utilizing multiple configurations. Equally, two classification techniques, namely Gaussian Mixture Model-Universal Background Model (**GMM-UBM**) and i-Vectors, were used to train and test the automatic detectors. The objective of this study was mainly twofold: firstly to evaluate the application of these techniques to a new scenario, analyzing their different degrees of freedom to establish a baseline to compare results with further studies; and secondly, to evaluate the influence of kinetic changes of instantaneous coefficients and the importance of the number of coefficients used to estimate the derivatives for the detection of **PD**. A large amount of trials were performed using a single parkinsonian and a single auxiliary database. The configuration leading to the best results was tested again with the other two remaining parkinsonian databases. Finally, cross-database trials were performed to validate the methodology at the optimum configuration. In these last trials, the correlation between the scores and **UPDRS** and **H-Y** scales was measured in order to analyze the convenience of employing these scores to evaluate the stage of the disease.

5.1.2 *Experimental set 2: forced Gaussian models*

In this second experiment, several approaches using **GMM** are proposed. In this experiment, a **FAM** was built to segment and label all the frames on the speech in the parkinsonian and **UBM** databases. The resulting phonetic labels were employed to build **GMM-UBM** models containing Gaussians which were specific for each phonetic label, yielding models able to compare all the phonetic units of the parkinsonian and control speakers more precisely. The different trials were performed in the three parkinsonian databases although, for some of these trials, only two databases could be used. Finally, a group of cross-database trials was performed to validate the optimal configurations.

The correlation between the scores and **UPDRS** and **H-Y** is measured in the obtained systems.

5.1.3 *Experimental set 3: allophonic distillation*

As several works in the literature point out, the consonants produced after a strong constriction of the articulators are usually more affected by **PD**. The idea behind the third experiment was to use only specific

segments of the speech, depending on the mode of articulation and the narrowing of the vocal tract to analyze their influence in the detection of PD. Again, the speech signals from the different databases are segmented and labeled (when possible) by using forced alignment techniques. However, in this case, the segments containing whole allophones were used to train and test the classifiers in the traditional manner, employing PLP as features. Forced alignment techniques identifying speech segments that were categorized into fricative, liquid, nasal, plosive and vowels, which are used separately to train and test GMM-UBM classifiers. Thus, the obtained models to detect PD were more influenced by the acoustic characteristics of these phonetic groups. The different trials were performed in the three parkinsonian databases although, for some of these trials, only two databases could be used. In this case, a fusion of scores coming from the models trained using different specific segments is carried out. Finally, a group of cross-database trials was performed to validate the optimal configurations.

The correlation between the scores and UPDRS and H-Y was measured in the systems too.

5.1.4 *Experimental set 4: acoustic landmarks*

In this fourth articulatory experiment, different types of points related to relevant articulatory moments, such as bursts, transitions between vowels and consonants or the beginning and end of glottal activity, among others, were used to identify relevant transition segments which were employed to detect PD, using an acoustic landmark distillation. This experiment is quite similar to Experiment 3 but, in this case, instead of using whole segments, only the transitions are used. To identify the acoustic landmarks and, thus, the transition segments, the methodology proposed by [197] was followed. In this case, new GMM-UBM classifiers were trained and tested using only transition segments characterized with PLP. Then, a fusion of scores was included to combine systems coming from different types of segments. All the points identified as landmarks were considered in the experiment although some of them were not true landmarks. The main reason behind this relies in the fact that if the algorithm detects a landmark, it can be caused by an articulatory movement or an abrupt change of energy in a certain frequency band. As the articulatory movements in parkinsonian patients can be, in some cases, subdivided, uncoordinated or repetitive, the use of all detected landmarks has a great potential for PD detection purposes. This methodology was applied to all of the parkinsonian databases. Finally, cross-database trials were performed to validate the methodology at the optimum configuration. Again, the correlation between the scores and UPDRS and H-Y scales was measured.

5.2 PHONATORY APPROACHES

A large portion of the approaches found in literature use jitter, shimmer and noise to detect PD but not many of them propose new features for this task. In the Experimental set 5, a new group of features, proposed by the author of this thesis were used in a GMM classification scheme to automatically discriminate patients from controls in three different databases. These features, containing information about noise, and frequency and amplitude perturbations on voice as well as tremor, were considered as appropriate for the proposed task.

This last experimental set can be considered supervised and language-independent. To finish the experiments, a fusion of scores between the phonatory approach and the articulatory approach yielding the optimal results was performed in order to combine the two aspects into a single approach.

An scheme of all of the experiments studied in this thesis and the final combination of aspects can be found in figure 5.1.

5.3 MATERIALS: SPEECH DATABASES

In this thesis, six databases are employed: GITA, Neurovoz, CzechPD, Albayzin, Vystadial-Czech and FisherSP. The first three contain different speech tasks from parkinsonian and control speakers. The mother tongue of the subjects in GITA and Neurovoz is the Spanish language, CzechPD database contain Czech speakers, as it could be inferred from its name. These three corpora are used to train and test the different classification models in the proposed approaches. On the other hand, Albayzin and Vystadial-Czech contain speech from native Spanish and Czech speakers respectively, and are used to train UBM models, when necessary in the proposed experiments. Finally, FisherSP is a speech database in Spanish language used to train the FAM employed in some experiments.

5.3.1 GITA database

GITA is a Colombian database presented in [127] containing a variety of speech tasks from 50 patients with PD and 50 age- and sex-matched control speakers whose native language is Spanish Colombian.

Four types of speech tasks from the GITA database are used in this thesis, comprising the DDK test (repetitions of the syllable sequence /pa-ta-ka/), a monologue, six fixed sentences and a sustained vowel /a:/. Average durations of these tasks are included in Table 5.3. The six fixed sentences are listed in Table 5.1, along with their translation into English while Table 5.2 shows the age, sex and severity statistics of subjects in the database.

Figure 5.1: Scheme of the overview of methodology employed in the thesis, including the four experiments on the articulatory approaches (a), the phonatory approaches, and the combination of the phonatory approach with one articulatory approach (b) where the articulatory part corresponds to the methodology providing the best results among the four articulatory experiments.

Table 5.1: Transcription of the six GITA fixed sentences (Spanish) and its translation to English

Sentence #	Spanish transcription / English translation
1	Luisa Rey compra el colchón duro que tanto le gusta / <i>Luisa Rey buys the hard mattress that she so much likes</i>
2	Los libros nuevos no caben en la mesa de la oficina / <i>The new books do not fit in the office desk</i>
3	Laura sube al tren que pasa / <i>Laura gets on the passing train</i>
4	Mi casa tiene tres cuartos / <i>My house has three rooms</i>
5	Omar, que vive cerca, trajo miel / <i>Omar, living nearby, brought honey</i>
6	Rosita Niño, que pinta bien, donó sus cuadros ayer / <i>Rosita Niño, who paints well, donated her paintings yesterday</i>

Table 5.2: Demographic statistics of GITA database. Ages are expressed in years.

	Female		Male	
	PD	HC	PD	HC
#Subjects	25	25	25	25
Age, average	60.7	61.4	61.6	60.5
Age range	49-75	49-76	33-81	31-86
UPDRS, average	37.6	-	37.7	-
H&Y, average	2.2	-	2.3	-
Years since diagnosis	12.6	-	8.9	-

Table 5.3: Average duration (s) of the speech tasks in GITA

Group	Sentence #						DDK	Monol.	Vowel /a:/
	1	2	3	4	5	6			
PD	4.0	3.4	2.2	1.9	2.5	4.3	4.7	46.0	3.3
Control	4.0	3.4	2.2	2.0	2.8	4.5	4.1	48.4	2.7
Both	4.0	3.4	2.2	2.0	2.7	4.4	4.4	47.2	3.0

Figure 5.2: UPDRS, H-Y and years since diagnosis histograms for the PD patients within the GITA database

The [UPDRS](#), [H-Y](#) and years since diagnosis distributions of the PD patients is portrayed in Figure 5.2.

The GITA database is sampled at 44.1 kHz and quantized with 16 bits. During the recordings, all patients were under pharmacological treatment, with no more than 3h since the last intake. The used microphone was a Shure , SM 63L, dynamic and omnidirectional. The mean Signal to Noise Ratio ([SNR](#)) measured¹ in this database is 29.3 dB.

5.3.2 Neurovoz database

The Neurovoz database is a new corpus containing 47 parkinsonian and 32 control speakers whose mother tongue is Spanish Castillian. This database was recorded in collaboration with the otorhinoaringology and neurology services of the Gregorio Marañón hospital in Madrid, Spain². The sub-corpus employed in this thesis contains a sustained vowel /a:/, a [DDK](#) test (/pa-ta-ka/), six fixed sentences and running speech during the description of the picture included in Figure 5.4. All of these tasks were produced at a comfortable speech [SPL](#). The average durations of the recordings are included in Table 5.6. Regarding the fixed sentences, these were first listened by the subjects from predefined recordings and then repeated, instead of read from a text docu-

¹ The mean [SNR](#) has been measured in the three parkinsonian databases employing five recordings per database, randomly selected from the control group utterances (monologue). To determine the mean noise level, at least 30 s of silences per database were extracted and to measure the mean signal level, 50 s of speech without silence were extracted per database too.

² The database was recorded within the project named "Evaluación multimodal de trastornos neurológicos mediante la caracterización de la voz, dinámica de los pliegues vocales y secuencias sacádicas (Neurovoz)", funded by the Ministry of Economy and Competitiveness of Spain, with reference TEC2012-38630-C04-01, in which the author of this thesis was granted with a predoctoral contract. The recruitment of patients and the recording of the corpus was approved by the Ethics Committee of the hospital.

Figure 5.3: UPDRS, H-Y and years since diagnosis histograms for the PD patients within the Neurovoz database

ment. Table 5.5 shows the age, sex and severity statistics of subjects in the database while Table 5.4 includes the transcription of the six fixed sentences and its translation to English.

The UPDRS, H-Y and years since diagnosis distributions of the PD patients is portrayed in Figure 5.3.

All of the patients were under pharmacological treatment and took the medication between 2 and 5 h before the speech and voice recording. These were assessed by a neurologist right before the recording session in order to assess their current neurological state. After this first examination, an otolaryngologist and a speech therapist assessed the patients' voice perceptually, and performed a clinical diagnosis through anamnesis and employing a visual examination of the vocal folds with a fiber-laryngoscope to discard organic pathologies³. A survey was conducted with the control group to assess their neurological state. A speech therapist assessed their voice perceptually and by means of a survey. Speakers with organic or neurological pathologies (other than PD) were discarded, as well as smokers and subjects with alcoholic addictions. Recordings of both groups were conducted in the same room with controlled acoustic characteristics.

³ 11 patients refused to do the fiber-laryngoscope exploration and were assessed only perceptually and through anamnesis

Figure 5.4: Image employed in monologue task in Neurovoz. Speakers were asked to describe the image during, at least, 30 s.

Table 5.4: Spanish transcription of the six Neurovoz fixed sentences and its translation to English

Sentence #	Spanish transcription / English translation
1	Cuando las barbas de tu vecino veas pelar, pon las tuyas a remojar / <i>When the beard of your neighbor you see peel, put yours to soak</i>
2	De la calle vendrá quien de tu casa te echará / <i>From outside will come who from your house will kick you out</i>
3	Cuando el diablo no sabe qué hacer, con el rabo mata moscas / <i>When the devil does not know what to do, kills flies with the tail</i>
4	La petaca blanca es mía / <i>The white flask is mine</i>
5	No pidas a quien pidió, ni sirvas a quien sirvió / <i>Do not beg the one who begged, nor serve the person who served</i>
6	El que a buen árbol se arrima, buena sombra le cobija / <i>To the one that comes to a good tree, good shade covers him</i>

Table 5.5: Demographic statistics of Neurovoz database. Ages are expressed in years.

	Female		Male	
	PD	HC	PD	HC
#Subjects	18	18	29	14
Age, average	70.9	68.4	71.9	66.6
Age range	59-86	58-83	41-88	55-77
UPDRS, average	18.2	-	7.4	-
H&Y, average	2.3	-	2.3	-
Years since diagnosis	6.6	-	7.4	-

Table 5.6: Average duration (s) of the speech tasks in Neurovoz.

Group	Sentence #						DDK	Monol.	Vowel /a:/
	1	2	3	4	5	6			
PD	4.6	3.3	3.8	2.3	3.6	3.9	10.7	29.8	3.6
Control	4.7	3.2	3.6	2.4	3.5	4.0	11.3	36.0	3.7
Both	4.7	3.3	3.7	2.3	3.5	3.9	10.9	33.8	3.7

The transducer used to record this database is an AKG C420 headset microphone which was coupled to a preamplifier with phantom power. The signal from the preamplifier was routed to a Soundblaster Live 24 bits sound card connected to a personal computer equipped with the software Medivox [198]. The sampling rate was 44.100 Hz, and the quantization, 16 bits. The mean SNR measured in this database is 24.3 dB

5.3.3 CzechPD database

The CzechPD database⁴ contains three speech tasks (a fixed text, a DDK task and a monologue) from 20 newly diagnosed and untreated parkinsonian speakers and 16 control speakers⁵ whose mother tongue is Czech. This database is described in detail in [95]. Unlike the other two parkinsonian databases, none of the patients CzechPD was under treatment. Table 5.7 shows the age, sex and severity statistics of subjects in the database while Table 5.8 includes the average durations of

⁴ The audios of this database were kindly shared by Dr. Jan Rusz in the frame of a collaborative work.

⁵ Only 14 speakers are available for DDK task.

Figure 5.5: UPDRS, H-Y and Years since diagnosis histograms for the PD patients within the CzechPD database

the speech tasks. The UPDRS, H-Y and years since diagnosis distributions of the PD patients is portrayed in Figure 5.5.

Since this corpus contains only a few speech tasks, it will be only employed in some tests with validation purposes.

Table 5.7: Demographic statistics of CzechPD database. Ages are expressed in years.

	Male	
	PD	HC
#Subjects	20	16
Age, average	61.0	61.8
Age range	34-83	36-80
UPDRS, average	17.9	-
H&Y, average	2.1	-
Years since diagnosis	2.38333	-

The sampling rate of the recordings of this database is 16 kHz and the quantification, 16 bits. The mean SNR measured in this database is 29.5 dB

5.3.4 Albayzin database

The Albayzin database [199] is a phonetically balanced dataset, sampled at 16 kHz and quantized with 16 bits, composed by a large amount of utterances in Spanish language and their transcriptions. For this thesis purposes, only the first subset from the five provided in the corpus is used with distinct objectives:

Table 5.8: Average duration (s) of the speech tasks in CzechPD. Ages are expressed in years.

Group	Fixed text	DDK	Monol.
PD	37.8	3.4	129.4
Control	38.2	3.2	118.5
Both	38.0	3.3	124.7

1. To create the **UBM** models, when needed, in the classification schemes employing GITA or Neurovoz as training databases (Experiments 1 to 4 in articulatory approaches).
2. To train the **FAM** to be used to automatically segment and label the fixed texts of GITA and Neurovoz.
3. To train a statistical model in the fourth articulatory experiment to detect acoustic landmarks.

5.3.5 *Vystadial-Czech database*

The Vystadial-Czech is a database in Czech language containing 18 h of recordings as described in [200]. The recording was performed via phone calls. However, the database is sampled at 16 kHz and 16 bits but a frequency analysis of the recordings reveals that there is no frequency information beyond 8 kHz, as it was expected since the recordings were made through the telephonic channel. The meta-data of the database includes the transcriptions of all the recordings although these have not been used in this thesis.

5.3.6 *FisherSP database*

The FisherSP (Fisher Spanish) has been created by the Linguistic Data Consortium to develop automatic speech recognizers in Spanish language. It contains around 163 hours of telephonic speech from 136 native Spanish language speakers from, at least, 20 countries but, mostly, from the USA. along with their transcriptions⁶. This database is sampled at 8 kHz and 16 bits. In this thesis, it is used to train the **FAM**.

5.4 GENERAL CONSIDERATIONS

Although the methodology followed for each experiment is different, there is a general scheme that is applied in most of the experimental sets carried out in this thesis, containing the following parts, as depicted in Figure 5.6:

⁶ Full description and transcriptions available at <https://catalog.ldc.upenn.edu/ldc2010t04>

Figure 5.6: General procedure followed in the thesis to analyze the different approaches.

1. Each signal was filtered and downsampled (if needed) to 16 kHz, to match the sampling frequency of the [UBM](#) databases or to other sampling frequency to limit the frequency information (specially in the phonatory approach).
2. Each signal, $s[t]$ was normalized to obtain $s_n[t]$ as follows:

$$s_n[t] = \frac{s[t]}{|\max\{s[t]\}|} \quad (5.1)$$

3. The signals are windowed using Hamming windows, using a certain frame length and frame shift, depending on the experiment.
4. After the windowing stage, each frame provided a feature vector, depending on the parameterization technique followed ([MFCC](#) or [PLP](#), for instance.)
5. The features were used to train and, subsequently, test a classification scheme ([GMM-UBM](#) or i-Vectors, for instance.)
6. The scores obtained from the previous stage were validated and a threshold was set to decide to which class ([PD](#) or control) corresponded the input utterance.
7. Finally, the results were analyzed using different types of merit figures such as accuracy, sensibility or sensitivity, among others, as well as graphic representations of the behavior of the systems.

Different types of validation of the proposed schemes were followed, like cross-validation (k-folds) or cross-database techniques, when possible. The cross database trials were carried out using the same type of speech tasks in the three databases, [DDK](#) task, since this is the only material that is phonetically equivalent in the three parkinsonian databases.

In some cases, this scheme is modified, including segmentation of the signals or other specific processing tasks. These particularities, are detailed for each experiment in the correspondent chapter.

In the parkinsonian databases, the sustained vowel was edited and does not include onset and offset segments. In the rest of the materials, silences at the beginning and end of the recordings were removed.

5.5 DISCUSSION

In this thesis, five different experiments are carried out employing articulatory and phonatory information from voice and speech to automatically detect PD. Three parkinsonian databases are utilized to train and test the PD automatic detectors, as well as other three auxiliary databases to create the UBM and the FAM, as it is summarized in Table 5.9.

Table 5.9: Summary of databases employed in this thesis.

Database	Brief description	Use in this thesis
GITA	Parkinsonian speech database in Spanish language containing sustained vowels, fixed text, DDK task and monologue.	Training and testing of models for the automatic detection of PD.
Neurovoz	Parkinsonian speech database in Spanish language containing sustained vowels, fixed text, DDK task and monologue.	Training and testing of models for the automatic detection of PD.
CzechPD	Parkinsonian speech database in Czech language containing fixed text, DDK task and monologue	Training and testing of models for the automatic detection of PD.
Albayzin	Speech database in Spanish language containing several hours of recordings with the corresponding transcriptions and other metadata.	Auxiliary database employed mainly to create UBM.
Vystadial-Czech	Speech database in Czech language containing several hours of recordings with the corresponding transcriptions and other metadata.	Auxiliary database employed mainly to create UBM.
FisherSP	Speech database in Spanish language containing several hours of recordings with the corresponding transcriptions and other metadata.	Auxiliary database employed mainly to create FAM.

As it has been mentioned previously, from the three parkinsonian speech databases used in this thesis, one, Neurovoz, has been recorded within the thesis period. The recording of a voice and speech database to study the influence of a pathology in the voice of the subjects must obey some rules to avoid an excessive influence of the channel in the audios, leading to biased studies or detectors in some cases. For example, some databases as the one used in this work, GITA, or other mentioned in Chapter 4 like the National Center for Voice and Speech Database included in works as [99], have been recorded in sound-treated booths. The fact that the booth is acoustically isolated does not mean that the acoustic conditioning, and therefore, the acoustic field inside the booth is the most adequate to record voice signals. In most of the cases, these cabins are designed to perform audiometries. In these cases, the acoustic isolation is crucial to allow the patient to listen exclusively to the audiometric signals using the headphones. But these booths commonly have multiple resonances, which produce a particular and pronounced impulse response of the room, leading to a remarkable and perceptible filtering of the voice signal that is recorded inside. This effect is even more problematic when the different groups (i.e., controls and patients) have been recorded in different booths or acoustic conditions. Considering the case in which the patients have been recorded in a small audiometry booth and the controls in a medium-size room, the audio recordings relative to the former will be affected by a pronounced filtering while the latter will not. This could bias many studies as, depending on the used features and methodologies, the channel could be detected as a differentiating factor between groups. Therefore, it is important to take into account all the acoustic elements which could affect to the recording, from the microphone to the acoustic environment. For these reasons, all of the speakers from Neurovoz included in this thesis were recorded in the same room with similar acoustic conditions and with the same equipment.

On the other hand, it is important to consider the most relevant phonetic differences between Castilian and Colombian Spanish, as reported in [201], that are:

1. /z/ and /c/ are pronounced in the same manner as /s/ in Colombian Spanish but differently in Castilian Spanish. The sentence "*el cielo es azul*" is pronounced as "el θjelo es aθul" in Castilian and "el sjelo es asul" in Colombian Spanish.
2. The phoneme⁷ /j/, velar in Castilian, turns into laryngeal /h/ in Colombian.
3. In Colombian, the consonants /b, d, g/ are pronounced as plosives when they are preceded by another consonant and as frica-

⁷ Some phonetic concepts such as "phoneme" or "plosive" are introduced in detail in Chapter 7 and 8

tives when the previous phoneme is a vowel. In Castillian these consonants are considered plosives only when are preceded by a nasal consonant (/m, n/).

Regarding these pronunciation differences between the speakers of the two parkinsonian databases in Spanish language, it was assumed that these will not influence significantly the results of the approaches proposed in this thesis. For instance, as it will be explained in the correspondent chapters, Albayzin has been used to create the [UBM](#) of the two databases indistinctly although there are accent differences between all of them: GITA contains Colombian accents, in Neurovoz the accents are mainly from the center of the Iberian Peninsula while Albayzin contains accents from several regions of the Iberian Peninsula. Something similar occurs with FisherSP, which includes Spanish accents from speakers all around the world. It was considered that for this thesis purposes, the differences in accents did not affect significantly the results of the procedures related to the automatic segmentation of the signals. In the same sense, it was assumed that the [DDK](#) tasks from the three parkinsonian databases are phonetically equivalent and the pronunciation in all the cases can be considered as similar.

Concerning the [UPDRS](#) and [H-Y](#) ratings of the patients of the three databases, these have been carried out in different health care centers or hospitals and by different neurologists. As these rating scales include subjective assessments of the symptoms and the stage of the disease for each patient, there can be an inter-database variability in the labeling. In the particular case of [UPDRS](#), all the values used in this thesis are global since the separate values of each of the parts of [UPDRS](#) (I, mentation, behaviour and mood; II, activities of daily living; III motor; IV complications.) were not available in any database. To this respect, the analysis of the correlation of the scores and the motor part of [UPDRS](#) is of interest.

With regard to the speech tasks, in the three parkinsonian databases the [DDK](#) task consists in the repetition of the syllable sequence /pa-ta-ka/ while the fixed texts are different in each database. While GITA and CzechPD include a monologue where the speakers describe what they did during the morning or any other day along the past week, in Neurovoz speakers were asked to describe an image. In this thesis, these running speech tasks are referred as "monologue".

Finally, although CzechPD contained the phonetic transcription of the fixed texts and Vystadial-Czech included the transcriptions of all the utterances, no automatic segmentations or exhaustive phonetic analysis were performed in these databases to avoid erroneous results derived from the inexperience in Czech language.

Part II

ARTICULATORY APPROACHES

EXPERIMENTAL SET 1: SPEAKER RECOGNITION TECHNOLOGIES

6.1 INTRODUCTION

Literature has demonstrated that PD influences the speech of patients in different manners. Derived from the analysis of a considerable amount of studies, it is reasonable to claim that an acoustic analysis of the speech, and particularly of articulation can help in the differential diagnosis of the disease.

As a matter of example, Figures 6.1 and 6.2 show the waveform and spectrogram of two patients, one newly diagnosed and other in an advanced stage and two respective age-matched controls during the repetition of the DDK sequence /pa-ta-ka/. At a first sight, the waveforms reveal several dissimilarities between patients and controls regarding articulation, like the differences in the dynamic margin and the abruptness between patients and controls. In the case of the newly diagnosed patient, most of the voice bars are nonexistent, and the burst in the articulation of /p/ tends to be shorter or substituted by a residual vibration of the vocal folds, transforming the spectrogram of this phoneme. In the case of the advanced stage patient, voice bars are proportionally shorter in comparison controls. In the same manner, his articulation is slower, which leads to differences in the spectrum of some segments respect to controls and is directly related to the velocity and acceleration of the articulators.

Therefore, and in order to perform an initial attempt to analyze the differences on speech between patients and controls using characterization and automatic classification schemes, the use of state-of-the-art speaker recognition technologies is proposed in the first experimental set. Thus, a thorough study about the influence of the different parameters and configurations of these techniques in a different environment, the detection of PD, is carried out. Principally, different combinations of acoustic material, parameterization and classification schemes are analyzed separately to identify the strengths and weaknesses of each one in PD detection. As it is depicted in Figure 6.3, speech materials can be a sustained vowel, a DDK task or two different sentences. Three families of features are used separately, namely, MFCC, Rasta-PLP and LPC as well as two classification schemes, GMM-UBM and *i-Vectors*. Each of these parts has some degrees of freedom such as frame length, number of feature coefficients or number of Gaussians among others. This experimental set proposes a large amount of combinations of these degrees of freedom, each one generating a different model which is tested and assessed in terms of accuracy for the detection of PD using the

(a) Idiopathic PD male speaker. Age: 58. UPDRS: 4. Span: 1,107 s

(b) Control male speaker. Age: 55. Span: 0,854 s

Figure 6.1: Waveforms and spectrograms of a newly diagnosed patient and a control speaker during a DDK task (three repetitions of /pa-ta-ka/). Obtained from Neurovoz database.

GITA database. In this way, it is possible to compare the performance of the classification methods and the suitability of the parameterizations and the speech tasks for the PD dysarthria detection. However, this experimental set considers trials using the different combinations separately not only to assess their convenience in PD dysarthria detection but to establish a baseline for comparison in future works.

It is important to remark that although literature shows some examples of the use of these techniques for the detection of PD [123], [129], [202], none of the works provides a deep study of them assessing thoughtfully the diverse degrees of freedom.

Additionally, a preliminary study of the kinetic changes is performed, since the speed and acceleration of the features which model the vocal tract can provide relevant information about articulation and its perturbations. In all of the cited combinations, this kinetic information is added to the feature vectors and the longitude of the speech segments employed to calculate them is evaluated in order to analyze its influence in the detection of PD. Particular attention has been paid to this degree of freedom which has not been thoroughly contemplated in the literature, independently of the features or classification scheme used, and on spite of the fact that fluctuations of articulation's velocity and acceleration are firmly associated to PD.

(a) Idiopathic PD male speaker. Age: 79. UPDRS: 23. Span: 2,820 s

(b) Control female speaker. Age: 73. Span: 1,182 s

Figure 6.2: Waveforms and spectrograms of an advanced-stage patient and a control speaker during a DDK task (three repetitions of /pa-ta-ka/). Obtained from Neurovoz database.

Once the analysis of all the mentioned factors is carried out, a final round of trials using other two parkinsonian databases (Neurovoz and CzechPD) to validate the results.

In short, the contribution of this experimental set is twofold, firstly to explore the possibilities of state-of-the-art speaker recognition techniques in the detection of [PD](#) establishing a baseline for future works. And secondly to assess the influence of kinetic changes in this detection.

6.2 THEORETICAL BACKGROUND

In this section, the main techniques and basis used in the methodology, i.e. the feature families, kinetic changes and classification schemes, are introduced along with a critical discussion of its use in automatic [PD](#) detection.

6.2.1 Parameterizations

A great number of short term parameterizations are employed in speaker recognition tasks but literature shows that the most recurrent features in this area are [LPC](#), [MFCC](#), and Rasta-[PLP](#). These three techniques are

Figure 6.3: Overview of methodology.

selected for this experimental set since all of them have been widely used for speaker and speech recognition tasks and have demonstrated their capabilities to characterize the vocal tract during the speech production. Each one provides a different characterization of the speech, allowing to extract some conclusions after the comparison of the different performances for the PD automatic detection with the proposed classification methods.

The first ones, LPC, were developed in the seventies [203] and are typically used to compress the audio signal, reducing its bit rate in communications. This type of feature extraction portrays largely the vocal tract and the radiation at mouth through the use of a time-variant all-poles filter which can describe the spectral envelope. LPC coefficients have not been tested in PD automatic detection.

MFCC are very well known in speech processing since the eighties [204] and can be considered as the standard speech parameterization approach. To obtain these coefficients, the spectral estimate of the signal is filtered in multiple bands and mapped in the mel scale. The logarithm of the resulting bands is transformed through the Discrete Cosine Transform (DCT) leading to the MFCC. Regarding voice and speech pathology detection, this parameterization has been successfully employed in a large number of studies with different purposes [205]–[208] including the detection of PD [123], [175], [202]. It is difficult to provide a direct interpretation of the information contained in these coefficients but it has been demonstrated that they can model the tract during articulation with some traces of glottal source-related features [209].

PLP coefficients were proposed in the nineties [210] as a cepstral domain representation of the vocal tract during articulation. To the author of this thesis knowledge, these coefficients have never been used for PD automatic detection. PLP analysis of speech is based on a linear prediction algorithm but in this case, the all-pole modeling is applied to an auditory spectrum, which is more consistent with human hearing. This auditory spectrum is obtained after warping the original spectrum into the Bark scale and applying a post-processing composed of two main steps: firstly, a pre-emphasis is applied simulating the equal-

loudness curve at 40 dB to consider the sensitivity of human hearing at different frequencies and secondly, a dynamic range compression is employed to compensate the non-linear relationship between the level of sound and its perceived loudness. This processing provides emphasis on the speech characteristics which are more relevant to the human hearing system and allows a low order in the subsequent all-pole model representing the vocal tract. Additionally, a Rasta filter can be included between the Bark filterbank and the auditory post-processing to obtain the Rasta-PLP. The idea of Rasta filtering is to eliminate all the components that do not contain phonetic information. To this end, a band-pass filtering is employed, alleviating the effect of convolutional noise introduced in the channel while smoothing some of the fast frame-to-frame spectral changes present in the short-term spectral estimate due to the analysis of artifacts [211]. The Rasta processing is a technique mainly used to mitigate the influence of noise in the codification, similar to others such cepstral mean subtraction. However, while cepstral mean subtraction compares the current analysis frame against the average of the whole utterance, Rasta uses a relatively short history of the signal on the order of several hundreds of milliseconds. Such a short history effectively enhances transitions between different speech segments and makes the result dependent on the previous short segment of speech such as phoneme or syllable [211].

A diagram explaining the procedure to calculate LPC, MFCC and PLP coefficients is introduced in Figure 6.4. A common trait of the three feature families is that they rely on the characterization of the resonance properties of the supralaryngeal vocal tract, which has been recognized as one of the most important sources of information in speech recognition applications [212]. However, the three groups of features are defined in different domains: LPC in the frequency and PLP and MFCC in the cepstral domain. It is convenient to remark that LPC do not use perceptual mechanisms to obtain coefficients while MFCC and PLP use perceptual approximations at some point but with different approaches. Thus, including the three parameterizations in the experimental set allows to compare the use of perceptual vs. non-perceptual and cepstral vs. non-cepstral parameterizations. In this sense, perceptual techniques can condense information while non-perceptual can provide information discarded by the other techniques.

With independence of the family of features used, it is possible to represent a vector $\mathbf{p}_n$ of F features calculated in a parameterization front-end for each time frame under analysis as:

$$\mathbf{p}_n = (p_1, \dots, p_i, \dots, p_F) \quad (6.1)$$

where n and i are the time frame and coefficient indexes respectively.

Figure 6.4: LPC, MFCC and PLP calculation procedure scheme.

6.2.2 Kinetic changes

Additionally to the three families of features, a characterization of the kinetic changes of each instantaneous coefficient is considered in this work in order to include variations of velocity and acceleration, which are destined to describe deviations in articulation.

One of the first attempts to include kinetic information in a speaker recognition system is reported in [213]. There, the author describes an automatic speaker verification system using cepstrum coefficients over short segments that are expanded by an orthonormal polynomial representation aiming to add dynamical information to the system. This expanded polynomial representation is also named *delta* (Δ) and describes the slope or velocity of the instantaneous coefficients. *Deltas* are typically calculated within a certain context segment of length $\tau_{derivative}$ (derivative segment length). In particular for the referenced work, *delta* features are defined over 90 ms segments containing 9 frames of 10 ms without overlapping since that interval length is found adequate for preserving transitional information between phonemes. In a further work by the same author [214], a speech recognizer relying on cepstral characteristics and *delta* features computed through polynomial regression is introduced. Using frames of 8 ms, the performance of the system is explored within different derivative segments varying from 3 to 11 frames for the *delta* coefficients computation. Results indicate that the inclusion of the *delta* significantly improved performance compared to just considering the instantaneous coefficients. Moreover, derivative segments of 72 ms -or 9 frames- are found to be the optimal value that derives effective regression coefficients. From this point on, literature has adopted the use of this kinetic changes in a wide selection of applications due to their complementarity with the information given by instantaneous coefficients. Likewise, several variations have been proposed, such as the inclusion of double-*delta* ($\Delta\Delta$) coefficients to characterise the curvature or acceleration.

Derivatives are calculated employing an anti-symmetric FIR filter and particularly for this work, the impulse response is obtained using the first order orthogonal polynomial as described in [214]:

$$D_{i,n} = \frac{\sum_{m=-n_0}^{n_0} C_{i,n}(m) \cdot m}{\sum_{m=-n_0}^{n_0} m^2} \quad (6.2)$$

where $D_{i,n}$ is the derivative at the n -th frame, $C_{i,n}(m)$ is a function defined in the interval $-n_0 \leq m \leq n_0$ containing $2 \cdot n_0 + 1$ coefficients centred at frame n and relative to the i^{th} coefficient of a certain feature family (for instance, the second MFCC). Thus, the number of frames used in the FIR filter and therefore the length of the signal segment employed in this process depends on n_0 . By varying this value it is possible to select the more appropriate derivative segment length or $\tau_{derivative}$ for specific purposes. Short values of $\tau_{derivative}$ might result in a noisy derivative sequence and too long values could smooth it, leading to loss of information. Figure 6.5 illustrates the relationship between $\tau_{derivative}$ and the number of coefficients and τ_{frame} used to calculate Δ . Figure 6.6 serves as an example of the use of different n_0 values in the derivative calculation where the higher n_0 , the smoother the derivative curve. Traditionally, works using derivatives do not explore the optimum n_0 and the number of FIR coefficients is not mentioned or is fixed to a value which is independent of the frame size. As $\tau_{derivative}$ is dependent of n_0 and τ_{frame} , using fixed n_0 when utilizing a range of different τ_{frame} produces the variation of $\tau_{derivative}$ according to the changes of the frame length. In other words, the use of short τ_{frame} generates shorter derivative segments than those obtained when using long τ_{frame} . This methodology could be unorthodox when trying to compare the performance of different τ_{frame} as these two parameters - τ_{frame} and $\tau_{derivative}$ - should be studied independently.

The first derivative - Δ , velocity- is obtained applying the expression 6.2 to the features vector and the second derivative - $\Delta\Delta$, acceleration- by using the same expression employing the first derivative array to extract $C_n(m)$. Therefore, after the parameterization of a frame, a new feature vector including derivatives can be represented as:

Figure 6.5: Scheme illustrating $\tau_{derivative}$ and its relationship with τ_{frame} and the number of coefficients used to calculate Δ for $n_0 = 2$.

$$\mathbf{x}_n = \{\mathbf{p}_n, \Delta, \Delta\Delta\} = (x_1, \dots, x_i, \dots, x_D) \quad (6.3)$$

where $D = 3 \cdot F$. Then, for an utterance composed of N frames, the cluster of feature vectors, $\mathbf{X}_u$ is:

$$\mathbf{X}_u = \{\mathbf{x}_1, \dots, \mathbf{x}_n, \dots, \mathbf{x}_N\} \quad (6.4)$$

Figure 6.6: First derivative of the fourth PLP coefficient ($F = 12$) calculated using two different values of n_0 during two /pa-ta-ka/ utterances of a control (upper) and a parkinsonian patient (down).

6.2.3 Classification schemes: GMM-UBM and *i*-Vectors

In the present experimental set two classification schemes, [GMM-UBM](#) and *i*-Vectors, are used. [GMM-UBM](#) have been employed in the last few years in speaker recognition and other speech-based tasks [195], having had great popularity until the appearance of the *i*-Vector approach

[215]. The *i-Vectors* technique has demonstrated to be successful in speaker recognition and in other speech-related tasks where some examples are diarization [216], language identification [217] or accent recognition [218]. The fundamentals of these techniques are explained next.

The general idea of a GMM model is to estimate the probability density function that characterizes the cluster of vectors of dimension D of a certain class, c , using a linear combination of G multivariate Gaussian components. The model can be used to calculate the likelihood of any new vector belonging to this class. This model is represented by $\Theta^c = \{\lambda_g^c, \mu_g^c, \Sigma_g^c\}_{g=1}^G$, where λ_g^c are the mixture weights $0 \leq \lambda_g^c \leq 1$, $\sum_{g=1}^G \lambda_g^c = 1$; μ_g^c are D -dimensional means and Σ_g^c , $D \times D$ -dimensional covariance matrices.

Thus, the likelihood that one feature vector $\mathbf{x}_n$ belongs to class c , given the model Θ^c is:

$$p(\mathbf{x}_n | \Theta^c) = \sum_{g=1}^G \lambda_g^c \mathcal{N}(\mathbf{x}_n; \mu_g^c, \Sigma_g^c) \quad (6.5)$$

where $\mathcal{N}(\cdot)$ represents Gaussian density functions.

The model Θ^c , with c indicating membership to a certain class $c : c \in \{\text{control}, PD\}$, is typically estimated using a given amount of speech frames belonging to class c following the maximum likelihood procedure using the iterative expectation-maximization algorithm [219]. When a large training set is supplied, this procedure provides a reliable estimation of the parameters of the model. However, the quality of the estimation is compromised when the training set is limited. It is possible to partially circumvent this by employing a larger auxiliary database which includes most of the speech characteristics under study, and which is modelled using a GMM procedure. This auxiliary model, $\Theta_{\text{ubm}} = \{\lambda_g^{\text{ubm}}, \mu_g^{\text{ubm}}, \Sigma_g^{\text{ubm}}\}_{g=1}^G$, is referred as *universal background model* (UBM) [195] and serves as a well-trained initialization, for which it is possible to adapt specific models using the provided -and more scarce- training data. The derived models, $\Theta_{\text{ubm}}^c = \{\lambda_g^{\text{ubm}}, \hat{\mu}_g^c, \Sigma_g^{\text{ubm}}\}_{g=1}^G$, are termed *GMM-UBM* and are expected to behave better than those developed using the training data directly. The adaptation from Θ_{ubm} to Θ_{ubm}^c is simplified adapting only the mean, μ_g^{ubm} to $\hat{\mu}_g^c$, using a Maximum a Posteriori (MAP) algorithm [195] and setting the covariance matrix and the weights to the values of the UBM, as it is typically done in speech related applications.

Therefore, the likelihood that the feature vector $\mathbf{x}_n$ belongs to class c , given the model Θ_{ubm}^c is obtained using expression (6.5) replacing Θ^c with Θ_{ubm}^c and their respective means and covariances.

The final decision about the membership of an input utterance to a certain class is taken by establishing a threshold over the score assigned to it, which is calculated by means of expression (6.6). The

threshold is typically set at the Equal Error Rate (EER) point determined using the training data.

$$\Lambda_{GMM-UBM} = \frac{1}{N} \log \prod_{n=1}^N \frac{p(\mathbf{x}_n | \Theta_{\text{ubm}}^{\text{PD}})}{p(\mathbf{x}_n | \Theta_{\text{ubm}}^{\text{control}})} \quad (6.6)$$

Another scheme to map the speakers into the two classes under study can be achieved rearranging the identity vector or *i-Vectors* approach [215]. The scheme employs the idea of GMM-UBM but using one model, $\Theta_{\text{ubm}}^u = \{\lambda_g^{\text{ubm}}, \hat{\mu}_g^u, \Sigma_g^{\text{ubm}}\}_{g=1}^G$, for each utterance instead for each class. Then, a transformation maps the means $\hat{\mu}_g^u$ to a $(G \cdot D)$ -dimensional vector $\mathbf{m}$ called *supervector* stacking mean vectors as follows:

$$\mathbf{m} = \{\hat{\mu}_1, \dots, \hat{\mu}_G\}$$

For a utterance $\mathbf{X}_u$, the utterance-specific-supervector $\mathbf{m}_u^c$, might be rearranged as:

$$\mathbf{m}_u^c = \mathbf{m}^{\text{ubm}} + T\mathbf{w}_u^c \quad (6.7)$$

where T is a rectangular matrix of low rank called *total variability matrix*; $\mathbf{w}_u^c$ is a random vector of dimension q (also called *i-Vector*) having a standard normal prior distribution $\mathcal{N}(0, 1)$; and $\mathbf{m}^{\text{ubm}}$ is the UBM mean supervector.

Since the *i-Vectors* scheme only models data in the low dimensional total variability space, compensation techniques are often applied to address the presence of variability factors affecting performance. One popular technique is the Gaussian Probability Linear Discriminant Analysis (GPLDA), which further decomposes the class and utterance dependent *i-Vectors* as [220]:

$$\mathbf{w}_u^c = \bar{\mathbf{w}} + \Phi \mathbf{y}^c + \epsilon_u \quad (6.8)$$

where Φ is an eigenvoice matrix, $\bar{\mathbf{w}}$ is the mean *i-Vector* class- and utterance-dependent, $\mathbf{y}^c$ is a vector of latent factors of dimension h and the residual ϵ_u contains the channel component and describes the within-speaker variability.

To compute the class membership of a certain test *i-Vector*, $\mathbf{w}_{\tilde{\mathbf{x}}}$, with a class c , a verification trial is performed to contrast the alternative hypothesis of the test *i-Vector* and the model having the same latent variables $\mathbf{y}^c$ ($\mathcal{H}_{\text{control}}$); or not ($\mathcal{H}_{\text{PD}}$). The model of the class c , $\bar{\mathbf{w}}^c$, is calculated as the mean *i-Vector* across the recordings belonging to c . In this manner, the score of the verification trial is as follows:

$$\Lambda_{i\text{-Vectors}} = \log \frac{p(\bar{\mathbf{w}}^c, \mathbf{w}_{\tilde{\mathbf{x}}} | \mathcal{H}_{\text{control}})}{p(\bar{\mathbf{w}}^c | \mathcal{H}_{\text{PD}}) p(\mathbf{w}_{\tilde{\mathbf{x}}} | \mathcal{H}_{\text{PD}})} \quad (6.9)$$

Figure 6.7: Example of DET and ROCCH-DET curves for two different sets of data.

These scores are calculated following the procedure proposed in [220]. The final decision is taken comparing the score with a threshold using the *BER* point¹ established during the training procedure.

6.2.4 ROCCH-DET curves

In this thesis, the Convex Hull around Receiver Operating Characteristic (ROCCH)-Detection Error Trade-off (DET) curves [221] are employed. These curves consist on the "lower left boundary of the convex hull around the discrete points of the Receiver Operating Characteristic (ROC)", as stated in [221], being, mainly, an interpolation of the discrete points of the ROC. This representation allows to determine more clearly the operating points and, although the interpolation could be considered as optimistic in some cases, authors of [221] demonstrate that the interpolated regions do not give lower minimum detection cost function for any operating point. This type of curve was proposed since the operating points of the minimum detection cost function do not yield over the horizontal or vertical lines of the ROC but in the interpolation between the vertices. Figure 6.7 shows an example of the traditional DET curves and the

6.3 EXPERIMENTAL SETUP

This section describes the databases and the methodology employed in this work.

¹ the *BER* is the score threshold or operating point for which the false rejection and the false acceptance rates coincide.

Table 6.1: Speech tasks and phonetic transcription (IPA) from GITA database. Sust. v. stands for sustained vowel

Speech tasks and phonetic transcription	
DDK	<i>pa-ta-ka</i>
	[pa 'ta ka]
Sentence 1	<i>Luisa Rey compra el colchón duro que tanto le gusta</i>
	[ˈlwisaˈreiˈkompra el kolˈtʃon ˈðuro ke ˈtanto le ˈɣusta]
Sentence 2	<i>Los libros nuevos no caben en la mesa de la oficina</i>
	[los ˈliβros ˈnweβos no ˈkaβen en la ˈmeʃa ðe la ofiˈʃina]
Sust. v.	<i>a</i>
	[a]

6.3.1 Acoustic material

Two databases are utilized in the first part of the experimental set: GITA and Albayzin. The purpose of these trials is to identify the optimum degrees of freedom such as the better τ_{frame} or n_0 , among others, in the automatic detection of PD using speaker recognition schemes. GITA database has been chosen since, among all the availables, it is well sex- and age- balanced and contains the same number of patients and controls. For this round of trials, four speech tasks from GITA are used, as detailed in Table 6.1. Most of the trials are performed for each one of these four groups of recordings separately. The tasks comprises the DDK utterances, two fixed sentences and a sustained vowel /a:/. The two fixed sentences listed in Table 6.1 have been selected from all the available resources since they contain a greater variety of articulation points. Tables A.1 and A.2 included in Appendix A contain their phonetic transcription using the International Phonetic Alphabet (IPA) and the place of articulation of the consonants (bilabial, labiodental, dental, interdental, alveolar, palatal or velar as illustrated in [222]). These places can be anterior (bilabial, labiodental, dental, interdental and alveolar), medium (palatal) or posterior (velar). In a similar manner, both tables indicate the presence of fast and slow glides (diphthong and hiatus respectively), and 2 or 3 joint consonants. Lastly, the sustained vowel is used for some limited tests in order to compare its performance against the other speech tasks.

In a second part of the experimental set, DDK, fixed texts and monologues from the three parkinsonian databases are used in a round of trials to validate the methodologies leading to best results in the first round.

6.3.2 Methodology

In the methodology of this first experimental set, two main parts are distinguished. In the first part, the initial assessment, the different degrees of freedom of the classification schemes are analyzed thoroughly in some speech tasks of the GITA database. After the analysis of this first round of trials, a second round using the three parkinsonian databases is performed to validate the conclusions obtained in the initial assessment by reproducing the best scenarios of the detection of PD with speaker recognition technologies under different circumstances.

6.3.2.1 Initial assessment

One of the main objectives of this experimental set was to test the state-of-the-art speaker recognition technologies in the detection of PD through speech. Nevertheless, the observation of the influence of the parameters involved in the detection, i.e. τ_{frame} , F , etc., can lead to additional conclusions. For instance, the analysis of the number of Finite Impulse Response (FIR) coefficients in the derivative calculation can point out towards finding the optimal $\tau_{derivative}$ in PD detection. In general lines, and as it is depicted in Figure 6.3, a parkinsonian speech database (GITA) was used as the input of a PD detector composed by a parameterization front-end and a classification system. These two main components have several degrees of freedom which were varied at each training-testing cycle using independently different speech tasks. It is important to remark that thousands of specific models or detectors were built and analyzed depending on the employed acoustic material, classification technique, family of features, τ_{frame} , $\tau_{derivative}$, G and other degrees of freedom specific of the classification technique. The joint analysis of these separated models allowed to evaluate the influence of the degrees of freedom in the PD automatic detection proposed in this work.

The degrees of freedom of the parameterization front-end were: family of features, F , τ_{frame} and n_0 . Before parameterization, all audio files from GITA database were filtered and resampled to 16 kHz. Signals from both databases were normalized and windowed using Hamming windows and τ_{frame} was varied from 10 to 40 ms at 5 ms steps. Family of features were MFCC, Rasta-PLP and LPC and the number of coefficients, F , varies in the range: $\{10, \dots, 20\}$ with steps of 2. After parameterization, derivatives were calculated using n_0 ranging from 1 to 7 avoiding $\tau_{derivative}$ values under 35 ms and over 80 ms. In that way, the use of a range of n_0 allowed to compare the influence of derivatives at different τ_{frame} maintaining the derivative segment lengths.

Respecting LPC and PLP, it is convenient to mention that among all the available techniques for the linear prediction algorithm, the autocorrelation method [203] was the one used in this experimental set.

```

\MAIN PSEUDOCODE STAGE 1
for each SpeechTask do:           // SpeechTask can be: DDK, Sentence 1, Sentence 2 or Sust. vowel*
  for  $\tau_{window}$  = 10 ms to 40 ms do:
    for  $n_0$  = 1 to 7 ** do:
      for each FeatureFamily do:   // FeatureFamily can be: MFCC, Rasta-PLP or LPC
        for F = 10 to 20 do:      // F = #features
          [TrainFEATURES, TestFEATURES] = CalculateFeatures {
            SpeechTask,  $\tau_{window}$ ,  $n_0$ , FeatureFamily, F }
          for G = 4 to 256 do:     // G = #gaussians
            model = train_GMM_UBM_Model { TrainFEATURES,
              Albayzin_DB, G }
            results = test_GMM_UBM_Model { TestFEATURES, model }
          return { results }
        end
      end
    end
  end
\ENDMAIN PSEUDOCODE STAGE 1

```

Figure 6.8: Pseudocode of Methodology. First stage. *Sustained vowel was used only in this stage. ** n_0 does not always range between 1 and 7; in some cases these values were restricted to limit $\tau_{derivative}$ between 35 and 80 ms

There were two stages in the first part of methodology in which most of the different degrees of freedom or parameters were varied, depending on the classification system employed. In the first one, only GMM-UBM classification schemes were used with G taking values from 4 to 256, in powers of 2. The pseudocode of this stage can be found in Figure 6.8 showing the different degrees of freedom involved, each one included in a `for` loop.

Albayzin database was used to generate the UBM, and each of the four speech tasks selected from the GITA database was employed separately for training and testing the GMM models following a k -folds cross-validation scheme, with $k=11$. An outline of this part of the methodology is shown in Fig. 6.9, including the ranges of all parameters. Among all the obtained results, those with better accuracy were considered the optimal although other indicators as sensitivity, specificity, Confidence Interval (CI), Area Under the Curve (AUC) and DET curve were also calculated in the assessment phase. CI was calculated in all cases as detailed in [223].

After this first round of modelling using GMM-UBM techniques, a second stage of training-testing with i -Vectors-GPLDA as classification system was performed. In this case, τ_{frame} and $\tau_{derivative}$ were fixed to the values providing best results in the GMM-UBM tests as it is detailed on the pseudocode of Figure 6.10 while the rest of the parameters from the parameterization front-end were varied in the same ranges as in the first stage (speech task excluding sustained vowel, family of features and F). In this scenario, the degrees of freedom of the classification system were the length of i -Vectors $-q$ - which takes up values $\{30, 50, 80\}$, dimension h of the GPLDA subsystem taking values $\{2, 6, 10, 14, 18\}$ and G varying in powers of 2 from 4 to 256. An outline of this part of the methodology is shown in Fig. 6.11. Equally, Albayzin database was used to create the UBM. An 11-folds cross-validation procedure was performed.

```

\MAIN PSEUDOCODE STAGE 2
for each SpeechTask do:           // SpeechTask can be: DDK, Sentence 1 or Sentence 2.
  for each FeatureFamily do:      // FeatureFamily can be: MFCC, Rasta-PLP or LPC.
    for F = 10 to 20 do:          // F = #features
      [TrainFEATURES, TestFEATURES] = CalculateFeatures { SpeechTask,
        Optimal  $\tau_{window}$ , Optimal  $n_0$ , FeatureFamily, F }
      for G = 4 to 256 do:        // G = #gaussians
        for q = 30 to 80 do:
          for h = 2 to 18 do:
            model = train_IVECTORS_GPLDA_Model { TrainFEATURES,
              Albayzin_DB, G, q, h }
            results = testGMMModel { TestFEATURES, model }
            return { results }
          endfor
        endfor
      endfor
    endfor
  endfor
\ENDMAIN PSEUDOCODE STAGE 2

```

Figure 6.10: Pseudocode of Methodology. Second stage.

Figure 6.9: GMM-UBM training methodology outline. Features can be MFCC, Rasta-PLP and LPC and derivative coefficients $\Delta + \Delta\Delta$.

In short, for the first stage (GMM-UBM classification scheme), the degrees of freedom were speech task, family of features, F , G , frame size and n_0 ($\tau_{derivative}$). Considering the best results obtained in this stage, the last two values were fixed to the optimum and were used in the second stage, (*i-Vectors* classification scheme) in which the degrees of freedom were the speech tasks (DDK and the two sentences), family of features, F , h , q and G .

The sustained vowel /a:/ was only used in GMM-UBM tests for comparison purposes. This acoustic material has poor articulation information which, in principle, is an important aspect for PD detection). Moreover, its use in combination with the *i-Vectors* scheme would not lead to more conclusions beyond those obtained in the GMM-UBM scheme. Although some works have proven that sustained vowels can carry information about the presence of PD, these utilize the onset and offset segments of the vowels during speech production and specific method-

ologies different to those included in this experimental set which are out of the scope of this work.

Figure 6.11: *i*-Vectors training methodology outline. Features can be MFCC, Rasta-PLP and LPC with derivative coefficients $\Delta + \Delta\Delta$.

In both stages, all the feature vectors of one specific speaker (X_u) were included in only one fold among all of the generated to perform the cross-validation. Each coefficient, x_i , from the feature vectors in the training folds was normalized in the interval $[-1, 1]$ while the weights obtained after this normalization process for each coefficient type, x_i , were applied to the test fold.

Table 6.2 shows the number of possible variations of each degree of freedom in the two stages, observing the two classification schemes and, for each one, the three parameterization families. For instance, as q can take values $\{30, 50, 80\}$, the number of variations was 3. The number of variations of n_0 range between 1 and 5 as the values that it can take depend on τ_{frame} . Each combination and fold leads to a new model. In the GMM-UBM scheme, an amount of 2856 models were trained and tested for each family of features and fold, while 1890 models were obtained in the *i*-Vectors stage. Therefore, the total number of models trained and tested in the whole experimental set was 156618, 14238 models per fold.

6.3.2.2 Validation and cross-database trials

Additionally, and in order to assess the influence of kinetic changes in global results, new tests were accomplished using only the configurations leading to the best results in the two stages (family of features, G, F, ...) but excluding derivatives from the feature vector. In the same manner, mutual information [224] of all features and kinetic changes respect to control and PD classes was calculated.

Table 6.2: Number of variations of each degree of freedom (Speech task, τ_{frame} , n_0 , F , G , q and h) in the two classification schemes and the three feature families

Number of variations for each degree of freedom				
Classification scheme	Degrees of freedom	Family of features		
		LPC	MFCC	Rasta-PLP
GMM-UBM (Stage 1)	Speech task	4	4	4
	τ_{frame}	7	7	7
	n_0	1 – 5	1 – 5	1 – 5
	# Coefficients (F)	6	6	6
	# Gaussians (G)	7	7	7
i-Vectors	Speech task	3	3	3
GPLDA (Stage 2)	τ_{frame}	1	1	1
	n_0	1	1	1
	# Coefficients (F)	6	6	6
	# Gaussians (G)	7	7	7
	q	3	3	3
	h	5	5	5

After these two stages of methodology, a last cluster of trials was performed to validate the results in other parkinsonian databases and with more speech materials. In particular, new [GMM-UBM](#) and *i-vectors* models were trained for the three parkinsonian databases using [DDK](#), fixed text and monologues separately. In this case, all the fixed text tasks of a database were employed to train a single model. In the case of CzechPD Database, the Vystadial-Czech database was used to build the [UBM](#). Moreover, and for comparison purposes, trials employing [DDK](#) tasks in the CzechPD database were repeated using Albayzin database. These results were expected to be similar in both cases as the pronunciation of /pa-ta-ka/ is practically the same in Czech and Spanish. In general, and in the rest of the experimental sets, the [DDK](#) task is using for inter-database comparisons as this is the only task which is common in the three parkinsonian databases. In these new trials, only the optimum feature family, τ_{frame} , $\tau_{derivative}$, h and q derived from the analysis of the two previous stages were used while the same ranges of F and G from other parts of the experimental set were analyzed again. Additionally, several cross-database trials were performed too, in which two parkinsonian databases were employed to train jointly a model while the remaining database was used to test this model. This process was repeated with all the possible combinations. In all of the trials, the Pearson's correlation (ρ) between the model scores

and UPDRS and H-Y ratings respectively were calculated, along with their correspondent p-values. In this case, DET curves are calculated, as in the initial assessment. However, and since some parkinsonian databases such as CzechPD have a limited amount of speakers, ROCCH-DET curves are calculated too.

6.4 RESULTS

For the sake of simplicity in the presentation, this section only includes results leading to best accuracy values and those showing a possible influence of the parameters in performance justified by the presence of PD.

6.4.1 Influence of τ_{frame} , n_0 and F in the first stage of methodology

In order to analyse the influence of τ_{frame} in the accuracy of the system, Figure 6.12 shows the best accuracy results as a function of τ_{frame} and family of features using a GMM-UBM classification scheme.

Figure 6.12: Best accuracies in the range $\{10, \dots, 40\}$ ms of τ_{frame} for the three different parameterizations, GMM-UBM classification techniques and using speech tasks: a) DDK, b) Sentence 1, c) Sentence 2, d) Sustained vowel. These values are obtained using different number of G , F and $\tau_{derivative}$ at each point.

In the same manner, to evaluate the impact of n_0 variation on accuracy with respect to the speech task materials and parameterizations,

Figure 6.13 shows the best results in terms of accuracy as a function of n_0 when using $\tau_{frame}=10$ ms, which is the optimum value in most of the cases as can be inferred from figure 6.12. In order to evaluate the influence of the number of coefficients of the FIR filter when using different frame lengths, Figure 6.14 depicts accuracy as a function of $\tau_{derivative}$ considering only Rasta-PLP+ $\Delta + \Delta\Delta$ and employing 10, 15 and 20 ms frame lengths.

Figure 6.13: Best accuracies using $\tau_{frame} = 10$ ms as a function of $\tau_{derivative}$ for the three different parameterizations using speech tasks: a) DDK, b) Sentence 1, c) Sentence 2. These values are obtained using different number of G and F at each point.

Additionally, Figure 6.15 shows the maximum accuracy obtained in the first stage as a function of the number of coefficients, considering $\tau_{frame} = 10$ ms.

Figure 6.14: Best accuracies using $\tau_{frame} = \{10, 15, 20\}$ ms as a function of $\tau_{derivative}$ for DDK, Sentence 1 and Sentence 2 using Rasta-PLP+ $\Delta + \Delta\Delta$ coefficients. These values are obtained using different number of G and F at each point.

Figure 6.15: Best accuracies using $\tau_{frame} = 10$ ms as a function of the number of coefficients for the three different parameterizations, GMM-UBM classification techniques and using speech tasks: a) DDK, b) Sentence 1, c) Sentence 2, d) Sustained vowel. These values are obtained using different number of G and $\tau_{derivative}$ at each point.

Figure 6.16: DET curves for MFCC, Rasta-PLP and LPC parameterizations $+\Delta + \Delta\Delta$ leading to the highest accuracy using the two classification techniques in Sentence 2 speech task tests.

6.4.2 GMM-UBM and *i-Vectors* global results

Best absolute accuracy in the first stage (86%) is obtained using $\tau_{frame}=15$ ms and $n_0 = 2$. Therefore, the two parameters are fixed to these optimum values for the tests in the second stage (*i-Vectors* scheme).

The best global accuracy results in the two stages of methodology are included in table 6.3. This table allows to compare the best outcomes depending on the speech task, parameterization and classification technique. Focusing on the sentence 2, which leads to the best absolute accuracy, DET curves for the three parameterizations and the two classification techniques are calculated (Figure 6.16). Figure 6.17 shows accuracy results as a function of G and F with the two different classification techniques and Rasta-PLP, which is the family of features which leads to best outcomes.

Table 6.3: Best accuracy results $\pm$ CI for the different acoustic materials using MFCC, Rasta-PLP and LPC with derivative coefficients $\Delta + \Delta\Delta$

Best accuracy results $\pm$ CI (%)						
Speech task	MFCC+ $\Delta + \Delta\Delta$		Rasta-PLP+ $\Delta + \Delta\Delta$		LPC+ $\Delta + \Delta\Delta$	
	GMM-UBM	<i>i-Vectors</i>	GMM-UBM	<i>i-Vectors</i>	GMM-UBM	<i>i-Vectors</i>
DDK	80 ± 8	80 ± 8	84 ± 7	82 ± 8	68 ± 9	74 ± 9
Sentence 1	77 ± 8	75 ± 8	83 ± 8	79 ± 8	65 ± 9	79 ± 8
Sentence 2	71 ± 10	77 ± 9	86 ± 7	87 ± 7	64 ± 10	71 ± 10
Sust. v.	74 ± 9	-	75 ± 8	-	60 ± 10	-

Finally, Table 6.4 shows the best global accuracy obtained for each speech task including their AUC, specificity and sensitivity and the specific configuration leading to these results. The values of AUC,

Figure 6.17: Best accuracies in GMM and *i-Vectors* tests using $\tau_{frame}=15$, $n_0=2$ with Rasta-PLP+ $\Delta + \Delta\Delta$ for Sentence 2 as a function of G and F . In the case of *i-Vectors*, $h = 50$ and $q = 14$. Best points of operation are marked with a red dot.

specificity and sensitivity are defined within the interval $[0, 1]$. Best global results are obtained using Sentence 2, Rasta-PLP+ $\Delta + \Delta\Delta$, *i-Vectors*, $q = 50$ and $h = 14$. It is possible to observe that Accuracy, AUC, Specificity and Sensitivity are the highest of all tests, having Specificity and Sensitivity similar values. Figure 6.18 shows DET curves for those results referred in the same table.

Figure 6.18: DET curves for the different speech tasks leading to the highest accuracies and the configurations detailed in Table 6.4.

Table 6.4: Best results for each acoustic material.

Best results										
Speech task	Accu. $\pm$ CI (%)	AUC	Spec.	Sens.	Feature family	τ_{frame} (ms)	n_0	Class. technique	G	F
DDK	84 ± 7	0,89	0,84	0,84	Rasta-PLP + Δ + $\Delta\Delta$	10	3	GMM-UBM	8	10
Sentence 1	83 ± 7	0,86	0,86	0,80	Rasta-PLP + Δ + $\Delta\Delta$	10	3	GMM-UBM	64	10
Sentence 2	87 ± 7	0,93	0,89	0,86	Rasta-PLP +Δ + $\Delta\Delta$	15	2	<i>i-Vectors</i>	4	12
Sust. v.	75 ± 8	0,83	0,70	0,80	Rasta-PLP + Δ + $\Delta\Delta$	10	3	GMM-UBM	256	10

Table 6.5: Results using Rasta-PLP and the configuration leading to best results with and without derivatives to analyze its influence in accuracy, AUC, specificity and sensitivity.

Feature family	Classification technique	Accu. $\pm$ CI (%)	AUC	Spec.	Sens.
Rasta-PLP + Δ + $\Delta\Delta$	GMM-UBM	86 ± 7	0,86	0,87	0,93
Rasta-PLP (without derivatives)	GMM-UBM	71 ± 9	0,68	0,74	0,78
Rasta-PLP + Δ + $\Delta\Delta$	<i>i-Vectors</i>	87 ± 7	0,93	0,89	0,86
Rasta-PLP (without derivatives)	<i>i-Vectors</i>	71 ± 9	0,74	0,68	0,79

Additionally, Figure 6.19 shows mutual information of features and derivatives leading to best results (Rasta-PLP and *i-Vectors* as referenced in Table 6.4). In the same way, Table 6.5 includes the results of this configuration with and without derivatives in the feature vector.

For the sake of simplicity, the influence of q and h in the results is not reported in detail since it has been considered that any conclusions regarding the performance of these parameters would be barely linked with the presence of PD. It is not possible to obtain any solid conclusions concerning q and h as the values providing best results are more influenced by the length of the database (i.e. number of speakers) and the length of the feature vector utilized rather than by the particularities of parkinsonian speech.

6.4.3 Validation and cross-database trials

Table 6.6 contains the results of evaluating the PD detection on the three available parkinsonian databases using Albayzin for the UBM and

Figure 6.19: Mutual information of Rasta-PLP (R-PLP) features with respect to the control/pathological classes

fixed text, [DDK](#) task and monologues as speaking tasks for GITA and Neurovoz, and the [DDK](#) task in CzechPD. To assess the use of this methodology in the CzechPD database with a [UBM](#) coming from Czech speech, Table 6.7 shows the results on the three available speaking tasks on the CzechPD database with Vystadial-Czech for the [UBM](#).

Tables 6.8 and 6.9 contain analogous results to the two previous, but in this case, employing an *i-Vectors-GPLDA* scheme for classification.

Finally, tables 6.10 and 6.11 include the results of cross-database trials employing [GMM-UBM](#) and *i-Vectors-GPLDA* classification schemes, respectively. In each trial, two parkinsonian databases have been used to train the models and the other has been used as test database.

Best global results for the k-folds cross-validation trials are in bold. For all the tables, Monol. stands for monologue.

Figures 6.20 and 6.21 show the [DET](#) and [ROCCH-DET](#) curves, respectively, of the best results referred in tables 6.6 and 6.8 for the three speech tasks while Figures 6.22 and 6.23 depict the [DET](#) and [ROCCH-acDET](#) curves for the three databases in the cross-database trials for the two different classification schemes.

6.5 DISCUSSION

In this first experimental set, the use of state-of-the-art speaker recognition technologies were analyzed for the detection of [PD](#). Firstly, a de-

Table 6.6: Results using GITA, Neurovoz and CzechPD in a GMM-UBM classification scheme, employing Albayzin for the UBM.

Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	F	G	UPDRS ρ	p-value	H-Y ρ	p-value
GITA	Fixed text	79 ± 8	0.86	0.73	0.84	14	256	0.12	0.4141	0.1	0.5136
	DDK	81 ± 8	0.88	0.82	0.8	18	256	0.07	0.6364	0.03	0.8589
	Monol.	80 ± 8	0.88	0.76	0.84	12	128	0.34	0.0166	0.16	0.2784
Neurovoz	Fixed text	86 ± 8	0.93	0.87	0.84	10	128	0.39	0.0193	0.12	0.4671
	DDK	79 ± 9	0.85	0.87	0.65	14	64	0.22	0.218	0.07	0.6627
	Monol.	72 ± 13	0.79	0.41	0.9	14	64	0.25	0.407	0.32	0.2854
CzechPD	DDK	88 ± 1	0.94	0.85	0.93	14	32	0.02	0.926	0.01	0.9601

Table 6.7: Results using CzechPD within a GMM-UBM classification scheme, employing Vystadial-Czech for the UBM.

Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	F	G	UPDRS ρ	p-value	H-Y ρ	p-value
CzechPD	Fixed text	80 ± 13	0.86	0.8	0.8	14	8	0.15	0.5218	0.18	0.4479
	DDK	91 ± 1	0.94	0.85	1	14	256	0.01	0.9814	-0.02	0.9476
	Monol.	83 ± 12	0.87	0.8	0.87	14	64	-0.06	0.811	0.11	0.6494

Table 6.8: Results using GITA, Neurovoz and CzechPD in a *i-Vectors*-GPLDA classification scheme, employing Albayzin for the UBM.

Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	F	G	UPDRS ρ	p-value	H-Y ρ	p-value
GITA	Fixed text	80 ± 8	0.85	0.82	0.78	18	128	0.18	0.2338	0.25	0.0925
	DDK	77 ± 8	0.8	0.74	0.8	12	64	-0.09	0.5515	-0.1	0.5021
	Monol.	77 ± 8	0.85	0.8	0.74	12	64	0.1	0.475	0.17	0.2469
Neurovoz	Fixed text	84 ± 8	0.93	0.77	0.94	12	64	0.12	0.499	0.12	0.48
	DDK	72 ± 10	0.86	0.72	0.73	14	256	0.13	0.4622	0.31	0.0545
	Monol.	79 ± 12	0.81	0.59	0.9	12	4	0.06	0.84	0.06	0.8576
CzechPD	DDK	76 ± 14	0.86	0.6	1	14	128	-0.31	0.1806	-0.21	0.3653

Table 6.9: Results using CzechPD within a *i-Vectors*-GPLDA classification scheme, employing Vystadial-Czech for the UBM.

Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	F	G	UPDRS ρ	p-value	H-Y ρ	p-value
CzechPD	Fixed text	66 ± 16	0.7	0.45	0.93	16	128	0.15	0.5273	0	0.9916
	DDK	68 ± 16	0.75	0.6	0.79	16	128	-0.19	0.4268	-0.05	0.8295
	Monol.	80 ± 13	0.92	0.7	0.93	20	32	-0.26	0.2753	-0.07	0.776

Table 6.10: Cross-database results using GITA, Neurovoz and CzechPD in a GMM-UBM classification scheme, employing Albayzin for the UBM. For every trial, two parkinsonian databases are used for training and the remaining, for testing.

Test database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	F	G	UPDRS ρ	p-value	H-Y ρ	p-value
GITA	DDK	73 ± 9	0.82	0.84	0.62	20	128	0.02	0.897	-0.22	0.1371
Neurovoz	DDK	75 ± 10	0.82	0.8	0.65	20	32	0.33	0.0543	0.33	0.0421
CzechPD	DDK	79 ± 14	0.91	1	0.5	12	4	0.21	0.3704	-0.01	0.9689

Table 6.11: Cross-database results using GITA, Neurovoz and CzechPD in a *i*-Vectors-GPLDA classification scheme, employing Albayzin for the UBM. For every trial, two parkinsonian databases are used for training and the remaining, for testing.

Test database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	F	G	UPDRS ρ	p-value	H-Y ρ	p-value
GITA	DDK	63 ± 9	0.67	0.64	0.62	16	4	-0.02	0.8823	-0.01	0.941
Neurovoz	DDK	67 ± 11	0.76	0.74	0.54	10	32	0.29	0.0942	0.13	0.432
CzechPD	DDK	76 ± 14	0.83	0.85	0.64	20	64	0.02	0.9397	-0.17	0.4755

tailed study of the main parameters and degrees of freedom involved in the whole detection scheme (such as family of features, frame length or classification technique, among others) in the GITA database using a sustained vowel, a DDK task and two fixed sentences. This first study leads to the optimal configurations in the detection of PD using speaker recognition technologies. Then, the methodologies are studied using the optimal configuration but employing the three parkinsonian and the two UBM databases. The exposed results allow to analyze the influence of parameterization, τ_{frame} , kinetic changes, speech task and classification schemes for the automatic PD detection through speech. The employed methodologies are the state-of-the-art techniques for speaker recognition but different optimum configuration parameters are expected for PD detection.

6.5.1 Influence of the family of features.

As it can be inferred from Tables 6.3 and 6.4, best results are obtained using Rasta-PLP+ $\Delta + \Delta\Delta$ coefficients, independently of the speech task and the classification techniques employed. In same cases, the performance of the MFCC family is similar to the one achieved with Rasta-PLP but providing a worse maximum accuracy. On the other hand, the differences in performance between LPC and the other two families of features are remarkable. Figures 6.12, 6.13 and 6.15 show that best results for LPC are usually lower than those obtained with Rasta-PLP and

Figure 6.20: DET curves for the three parkinsonian databases and the two classification schemes. Albayzin is used as UBM for Neurovoz and GITA trials and Vystadial-Czech for CzechPD trials.

MFCC and in many cases are at 20 absolute points under the best Rasta-PLP accuracies.

Going into detail to the Rasta-PLP parameterization, to evaluate the influence of Rasta filtering in the whole process, some prospective tests are performed using PLP with and without Rasta filtering to compare its influence in PD detection. To this aim, the GMM-UBM methodology depicted in Figure 6.9 is repeated limiting $n_0 = 4$ using DDK speech tasks and a maximum τ_{frame} of 30 ms. Figure 6.24 shows the best accuracy vs. frame length obtained in the prospective tests to analyse the use of Rasta filter, while the best numeric results are included in table 6.12. As it can be inferred from these outcomes, Rasta filtering does not provide remarkable improvements to PLP suggesting that this filtering does not enhance the capabilities of PLP for the detection of PD in this experimental set.

Figure 6.21: ROCCH-DET curves for the three parkinsonian databases and the two classification schemes. Albayzin is used as UBM for Neurovoz and GITA trials and Vystadial-Czech for CzechPD trials.

6.5.2 Influence of τ_{frame} and $\tau_{derivative}$.

With reference to τ_{frame} , in general, best results are obtained with the shortest frames, 10 and 15 ms, which is consistent with the adoption of quasi-stationary conditions. Something similar occurs with the derivative segment length in which a high $\tau_{derivative}$ could blur the velocity and acceleration tracking. Figure 6.13 shows the accuracy obtained with the three parameterizations as a function of $\tau_{derivative}$ for $\tau_{frame} = 10$ ms. In this scenario, best results are obtained using $n_0 = 3$, which corresponds with the use of 7 coefficients in the FIR filter and a $\tau_{derivative}$ of 40 ms. Using different τ_{frame} , the optimum $\tau_{derivative}$ is similar, as it can be inferred from Figure 6.14 in which the derivative segment lengths providing best results are most of the times around 40 ms. This suggests that the analysis of velocity changes in parkinsonian speech characterization should be performed within these ranges. This length allows more resolution to characterize imprecise articulation co-

Figure 6.22: DET curves for the three parkinsonian databases and the two classification schemes. Albayzin is used as UBM in all the cases. Only DDK task is employed

Figure 6.23: ROCCH-DET curves for the three parkinsonian databases and the two classification schemes. Albayzin is used as UBM in all the cases. Only DDK task is employed

Table 6.12: Results using PLP feature family with and without Rasta filter in order to its influence in accuracy, AUC, specificity and sensitivity.

Parameterization	τ_{frame} (ms)	Accuracy $\pm$ CI (%)	AUC	Spec.	Sens.
PLP+ Δ + $\Delta\Delta$	15	81 $\pm$ 8	0.87	0.82	0.80
Rasta-PLP+ Δ + $\Delta\Delta$	10	82 $\pm$ 8	0.87	0.82	0.82

Figure 6.24: Best accuracies for PLP parameterization with and without Rasta filtering, using DDK speech task and GMM-UBM classification technique. τ_{frame} is varied in the range $\{10, \dots, 30\}$ ms

ordination than larger segments, which is coherent with the dysarthria features found in [122], [225].

6.5.3 Influence of kinetic changes.

A supplementary analysis of the relevance of kinetic changes in the group of features is performed. As derivatives characterize the velocity and acceleration of vocal tract movements for certain n_0 values, the analysis of its relevance can help disclosing the importance of articulation in the studied detectors. Firstly, Figure 6.19 shows mutual information of all the coefficients leading to the best results (as detailed in Table 6.4, Sentence 2) and the speaker's condition. The mutual information relative to kinetic coefficients is significant with regard to the mutual information of static Rasta-PLP coefficients. Besides, several tests are repeated introducing some variations to assess the influence of these derivatives in the accuracy. These new tests are performed employing 12 Rasta-PLP coefficients without derivatives, $\tau_{frame}=15$ ms, and the two classification schemes using values of $\{4, 8, 16, 32\}$ for G . In the case of i -Vectors, q and h are varied in the same range specified in the methodology of the experimental set. After removing derivatives, tests give a maximum of 71% of accuracy in both cases as shown in Table 6.5, confirming the influence of the kinetic changes and the importance of articulation.

6.5.4 *Influence of speech tasks.*

Regarding the speech tasks and taking into account the overlapping margins in accuracy caused by the CI, the results suggest that Sentence 2 leads to better accuracies, specially when parameterizing the signal following the Rasta-PLP technique.

In that sense, the models trained with the sustained vowel are the least efficient among those obtained in the experimental set. These results support the importance of articulation for PD automatic detection with speaker recognition techniques. As it is introduced in Chapter 4, the works centered in phonatory aspects providing good results make use of the phonation of different vowels or extract vowel segments from running speech, which contain certain articulatory information. In the present case none of these circumstances occur, which explains why accuracy does not exceed 75% with the sustained vowel task as shown in Figure 6.12 d). However, works like [82], [122] suggest that phonation by itself can provide information about tremor or noise attributable to PD, but to take advantage of these circumstances, the use of features characterizing glottal source is necessary. Going into detail to Figure 6.12 d), results with LPC show poor accuracy as expected since this family of features does not contain relevant information about the glottal source. Better outcomes are achieved using MFCC and Rasta-PLP coefficients which suggest that both families can characterize in some manner the glottal source, and that could justify the relatively good results that they provide (75% accuracy) using only one sustained vowel.

When comparing the speech tasks in the three different databases, fixed texts usually produce better results in the analyzed classification schemes. One of the main reasons for this behavior is that the GMM are limited to a certain group of allophones and/or segments, which are repeated in the testing utterances too. Being PD a condition affecting movements, which may influence distinctly different types of articulatory movements, the use of text-dependent systems will allow to better compare specific movements in a new speaker with the two possible classes (through the models) as the same type of movements are compared (those used in training and testing). Thus, the use of monologues as speech task with the same purposes, employing the analyzed schemes will provide a more vague comparison and will need more training data to create the models, in order to better represent the articulatory properties of the speakers which could appear on testing and larger testing data per speaker, to compute the likelihood respect to the models with a wider variability.

6.5.5 *Influence of classification techniques.*

Regarding the classification techniques, GMM-UBM methods give generally the best results with all acoustic materials but Sentence 2, which

provides the best outcomes combined with *i-Vectors* in the initial assessment. In general, the differences of best accuracy results comparing both schemes are small. Moreover, although the *i-Vectors*-GPLDA methodology is more sophisticated and usually produces better performance for speaker recognition, it is more oriented to use larger UBM and training databases, containing more variability than the ones used in this experimental set. Moreover, it is possible to observe in Figure 6.17 that the highest accuracy results are more spread in the GMM case (left) with respect to the *i-Vectors* best results (right), which are concentrated into a small region indicating that the first technique could generalize better. Focusing on the validation trials, all GMM-UBM schemes outperforms in terms of accuracy and AUC in most of the cases in the three databases. These differences are even more pronounced in the cross-database trials.

6.5.6 UPDRS and H-Y assessment

In general, the correlation between the scores and the UPDRS and H-Y ratings is non significant. Only significant values are obtained with monologues in the GITA database employing GMM-UBM, being the highest correlation 0.34 respect to UPDRS. Considering that it is plausible to think that there is a certain correlation between the scores, representing the likelihood of a speaker to suffer from PD, and the patient's UPDRS rating, this correlation could be stronger with the motor part of the rating, UPDRS part II, as the classifiers are using motor information (speech articulation) to obtain the scores. However, in this thesis, only the global UPDRS was available and this assumption cannot be verified. This can explain the limited correlations, as the global ratings include other non-motor aspects with low correlation with the motor dysfunctions.

6.5.7 Other considerations.

It is known that the three families of features provide information about the vocal tract and the added $\Delta + \Delta\Delta$ coefficients can capture these changes during articulation. Nevertheless, LPC parameterizes the vocal tract in the frequency domain, whereas PLP and MFCC provide the information in the cepstral domain and use perceptual modelling. These two facts seem to be relevant in the speech processing for the detection of PD. Additionally, the presence of glottal source information in MFCC and, possibly, in Rasta-PLP could contribute to these good results. Taking into account the curves depicted in Figure 6.12 d), in which results are similar in MFCC and Rasta-PLP, there could be a causality relationship between the possible existence of glottal source information in PLP (as in MFCC) and the differences in the results ob-

tained using the sustained vowel respect to [LPC](#), but this existence cannot be completely confirmed with the performed tests.

On the other hand, the optimum configuration provides values of 87% in accuracy. This value is similar to most of those reported in the state-of-the-art of [PD](#) detection. For instance, in [\[122\]](#), values up to 88% of accuracy are obtained. Other works as [\[95\]](#) obtain values up to 80% using vowels containing articulatory information extracted from monologues, the same value reached in [\[175\]](#) when combining phonatory, prosodic and articulatory features for [PD](#) automatic detection. Nevertheless, results are difficult to compare when different databases are employed, taking into account that in most of the cases they differ in language, sex and age distribution, speaker tasks and number of speakers. Moreover, there are not too many works analyzing automatic detectors of [PD](#) based only on voice or speech. Some works provide hybrid methods combining speech with other sources such as posture and gait [\[226\]](#) and others use very small and unbalanced databases [\[103\]](#), [\[227\]](#). This makes unfeasible to perform too many comparisons with the obtained results at the moment.

Also, the literature provides preliminary attempts of using automatic speaker recognition procedures such as *i-Vectors* and [GMM-UBM](#) to detect or assess [PD](#) [\[123\]](#), [\[129\]](#), [\[202\]](#). Some of these schemes employ multiple features but without a thoughtful consideration of the degrees of freedom or a comparison of results between different methodologies. For instance, [\[123\]](#) uses [MFCC](#) and [GMM-UBM](#) techniques with running speech to detect [PD](#) but employs the same database for [UBM](#) modelling and training-testing, which could bias the results. Moreover, the work does not analyse the different parameterization and classification ranges or degrees of freedom. In this sense, the study of the optimum parameters for speaker recognition has been addressed in multiple works and these have been reconsidered in the automatic [PD](#) detection scenario in this work.

Regarding the [CI](#) obtained on this experimental set, in most of the cases it is relatively high in comparison to those obtained in state-of-the-art speaker recognition works in which typical values are around 0.5%. These [CI](#) values could be enhanced in two manners. The first one could be improving the accuracy results although, as mentioned, the false rejection could have a theoretical limit. The second one can be achieved using more speakers, but obtaining large databases of speech from [PD](#) patients implies an important cost. That is why most of pathological voice (or speech) databases are usually small in comparison to speaker or speech detection aimed databases. In general, the high values of the [CI](#) limit the conclusions that can be obtained after observing the results since there is a clear overlapping between the accuracy values when considering the [CI](#). Therefore, when concluding that fixed texts provide better results, for instance, this conclusion is based upon a trend observed in the accuracy and [AUC](#) but it is not possible to claim such conclusion with total certainty.

With respect to the use of **ROCCH-DET** curves, these provide more clear interpretation of the trends of the different classifiers and allow to compare the different approaches more easily in the cases in which the traditional **DET** curves are quite "steppy". Therefore, and although these curves can be considered optimistic from certain points of view, **ROCCH-DET** curves will be employed in the thesis in the following chapters for the sake of comprehensibility.

Lastly, it is convenient to contemplate that, as it is considered in [150], some symptoms as depression can influence the speech rate and therefore articulation. In the recording of the GITA database the presence of depression in patients is not assessed and, thus, this could bias the results.

While best results, 87% of accuracy, are obtained using Rasta-**PLP** parameterization and fixed sentences, none of the classification techniques seems to stand out respect to the other. However, the potential of *i-Vectors* is not exploited totally since this technique is more aimed to work with larger and more complex databases, including more variability. An adaptation of this methodology for small databases could be advisable in future works. As results suggest that the variability between the two studied classes is small, techniques to reduce intra-class variance should be applied in the future to improve classification results.

The tuning of the $\tau_{derivative}$ for velocity calculation provides a margin of gain of more than 5 absolute points in classification, concluding that the optimum value is around 40 ms. In the future, the study of kinetic changes and particularly the tuning of acceleration must be addressed and its influence in the global detector evaluated. New features based on velocity and acceleration related to articulation should be considered.

Additionally, although the signs of **PD** appear on each patient in a particular manner, it could be advisable to analyse its influence on specific types of phonemes in order to identify articulation places or movements more likely to include perturbations caused by the disease. Moreover, as automatic detection of **PD** is achieved essentially by means of the detection of parkinsonian dysarthria, further studies should analyse the differences between the influence of this type of dysarthria and the one caused by non-parkinsonian neurological disorders.

Finally, it is possible to conclude that the state-of-the-art techniques for speaker recognition provide good accuracy results in the automatic detection of **PD** and can be considered as a baseline in future works. Results suggest that Rasta-**PLP** characterization provides the best accuracy compared to **MFCC** and **LPC** using short τ_{frame} (10-15 ms) and fixed sentences as acoustic material. The kinetic changes have to be taken into account in the future as their relevance has been proven in the automatic detection of **PD** patients when using 40-50 ms ranges for $\tau_{derivative}$.

6.6 SUMMARY

In this chapter, two classification schemes (GMM-UBM and *i-Vectors*-GPLDA) are employed separately with several parameterization techniques, namely [PLP](#), [MFCC](#) and [LPC](#). Additionally, the influence of the kinetic changes, described by their derivatives, is analyzed.

These results are comparable to those obtained in other works employing speech for [PD](#) detection and confirm that the selected speaker recognition techniques are a solid baseline to compare with further experimental sets. Results suggest that Rasta-[PLP](#) is the most reliable parameterization for the proposed task among all the tested feature families while [GMM-UBM](#) provides better results and generalization than *i-Vectors*-[GPLDA](#). Additionally, results confirm that the use kinetic changes improve substantially the accuracy in [PD](#) automatic detection systems.

Cross-validation trials (k-folds) provide accuracies between 81% and 91%, with [AUC](#) between 0.88 and 0.94 depending on the employed database, while cross-database trials provide accuracies between 73% and 79% with [AUC](#) between 0.82 and 0.91.

EXPERIMENTAL SET 2: FORCED GAUSSIANS

7.1 INTRODUCTION

After the use of speaker recognition technologies which provide good results in the detection and assessment of PD, a new approach, considering the particularities of articulatory movements in parkinsonian speakers, is proposed here. To this respect, works like [48], [105]–[108], [228], among many others, point out that PD affects differently to the various articulatory movements.

To this respect, Figures 7.1 and 7.2 illustrate two examples of the possible differences in articulation and phonation between patients and controls. In the first case, Figure 7.1 shows the differences in waveform and spectrogram of the phoneme /e/ between a newly diagnosed patient and a control during the pronunciation of the syllable /pe/ within the word "petaca" in Spanish. In the case of the patient, the vowel presents more irregularities, which is reflected in the spectrogram as an absence of well defined formants. In the second example, both speakers pronounce the word "es" within the Spanish sentence "La petaca blanca es mía". In the case of the patient it is observed that the articulation of the phoneme /s/ contrasts with the analogous of the control since, in the first case, it becomes almost a continuation of the phoneme /e/ as it is reflected in the waveform and in the spectrogram where the /e/ formants do not disappear. This effect may be caused due to an incoordination in the use of the glottal source which leads to continuous vibration of vocal folds even during the articulation of the /s/ consonant, where an interruption of the phonation was expected.

Therefore, and in order to perform a more efficient classification of the patients and controls in a binary detection system, in this experimental set the different phonemes were used separately in the classifiers, providing systems allowing to discriminate between the two types of speakers based on statistical models of these phonemes.

7.2 THEORETICAL BACKGROUND

7.2.1 *Phonetic considerations*

At this point, it is important to discern between two concepts: phonemes and allophones. A phoneme is a phonetic unit which cannot be decomposed in minor unities and which can distinguish one word from

(a) Idiopathic PD female speaker. Age: 59. UPDRS: 9. Span: 61 ms

(b) Control female speaker. Age: 59. Span: 65 ms

Figure 7.1: Waveforms and spectrograms of a parkinsonian (newly diagnosed) and a control speaker pronouncing the syllable /pe/. Obtained from Neurovoz database.

other. For instance, in Spanish¹ the words "pez" and "tez" contain three phonemes each and both are distinguished thanks to their first phoneme. The term allophone is referred to each of the pronunciation variants of a certain phoneme, usually, as a function of this phoneme in a word or syllable and depending on the adjacent phonemes. For instance, the phoneme /g/ has the allophone g the word "gusta" and y in "disgusta".

7.2.2 Speech forced alignment

Speech forced alignment techniques [196] are used to detect allophones in an utterance, knowing its transcription (and therefore, the sequence of phonemes of the speech signal), resulting in the automatic segmentation of the signal into separated allophones and the labeling of each of these segments with their correspondent phoneme label. This technique can be considered a supervised technique, since, as a difference

¹ All the examples here are included in Spanish as the main databases used in the experimental sets are in Spanish and the categorization of allophones and phonetic rules, in general, are referred to this language.

(a) Idiopathic PD female speaker. Age: 85. UPDRS: 47. Span: 180 ms

(b) Control female speaker. Age: 83. Span: 180 ms

Figure 7.2: Waveforms and spectrograms of a parkinsonian (in an advanced stage) and a control speaker pronouncing the word "es". Obtained from Neurovoz database.

to automatic speech recognition techniques based on [DNN](#) or Hidden Markov Models ([HMM](#)), in this case the system must know the transcription of the utterance analyzed beforehand. An example of forced alignment obtained with a model trained using Kaldi development kit [229] and the utterance "La petaca blanca" is shown in Figure 7.3.

In this experimental set, a [FAM](#) is obtained with Kaldi following a [GMM/HMM](#) architecture, similar to the one explained in [230]. In general lines, the training of this [FAM](#) follows these steps:

1. All the speech signals are characterized using [MFCC](#)+ Δ + $\Delta\Delta$ features and metadata such as lexicon and dictionaries auxiliary files are generated.
2. Using these features, of a [GMM](#)-based monophone model is trained, taking into account all the possible allophones contained in the training corpus². This first model consists in a group of [GMM](#) models, representing each of the singular allophones or monophones. It is obtained after several iterations of model training and aligning³ the utterances. In the first iteration, the utterance

² Notice that the corpus must include the transcription files of all the utterances

³ Notice that alignment is referred to segmentation and labeling of the speech at the same time. The segmentation and the labeling are used in each iteration to train the

Figure 7.3: Example of forced alignment labeling obtained with Kaldi, including waveform, spectrogram, phonemes and transcription

is subdivided into a number of chunks equal to the number of monophones of the utterance, obtained from the transcription. Each chunk has the same time length in the first iteration but after a re-aligning using [HMM](#) and Viterbi algorithms, new segments are obtained and used to train new [GMM](#) models. This process is iterated for a certain number of times or until there is not substantial improvement in this alignment and it is considered that there is a convergence.

3. Using this monophone first basic alignment as a basis, a new model is created following the same procedure but considering adjacent allophones. Therefore, a triphone model is obtained where there is a clustering [GMM](#) representing any possible combination of monophones, i.e. triphones. This model is optimized after several iterations, as in the previous step.
4. Once this new model is obtained and all the signals are aligned employing triphones, all the features are transformed using the Linear Discriminant Analysis ([LDA](#)) + Maximum Likelihood Linear Transform ([MLLT](#)) projection. These are employed in a new sequence of training-alignment iterations considering triphones.
5. Finally, a Feature space Maximum Likelihood Linear Regression ([fMLLR](#)) transformation of the features is used for speaker adaptation in order to obtain a speaker invariant model during the last iterations of the triphone training-alignment process.

This model is applied to the utterances of the fixed texts and sentences of the databases, characterized by means of [MFCC](#) + Δ + $\Delta\Delta$ to automatically segment and label all these speech recordings. It is important to remark that the [MFCC](#) characterization is used only to train the [FAM](#) and to obtain the final labels but in this experimental set it is not used directly to detect [PD](#). In the [FAM](#) model obtained following a Kaldi recipe, these labels have the format ζ_T , where ζ is the phonetic

---

new [GMM](#) models, as the segmentation indicates which part of the utterance will be used to train a specific [GMM](#), represented by the labeling.

label and T is a subscript indicating the position of the phoneme referred by the label within the word in the transcription. The index ξ represents, in most of the cases, the phoneme that the label is representing (i. e., /b/, /a/, or /n/), without making any distinction between the different types of possible allophones that can be associated for a single phoneme (only the plosive-fricative allophones of the phonemes /p/, /d/ and /g/ are considered). On the other hand, T can be represented by the symbols $\{B, I, E, S\}$ where B indicates beginning, I , intermediate, E , end and S , single or separated.

7.2.3 Forced Gaussian Mixture Model

One of the first techniques proposed to improve the **GMM-UBM** scheme is the Forced Gaussian Mixture Model (**fGMM**) technique. This new scheme consists on the training of a **GMM** classifier in which the number of Gaussian mixtures is determined by the number of phonemes included in the training database and the expectation-maximization process is replaced by a direct assignment of all the training frames obtained from a certain phoneme, to their correspondent Gaussian mixture.

Therefore, the idea is to obtain **GMMs** which allow to compare more precisely all the different phonetic segments of the two classes.

For a certain training database containing speakers from two classes, $\mathbf{Y}_c$, defined in 7.1, is the set of utterances of class c , composed of U_c utterances $\mathbf{X}_u$:

$$\mathbf{Y}_c = \{\mathbf{X}_1, \dots, \mathbf{X}_u, \dots, \mathbf{X}_{U_c}\} \quad (7.1)$$

Then, for an utterance containing N frames, the cluster of feature vectors, $\mathbf{X}_u$ is:

$$\mathbf{X}_u = \{\mathbf{x}_1, \dots, \mathbf{x}_n, \dots, \mathbf{x}_N\} \quad (7.2)$$

where the vectors have dimension of size D . For any utterance, let Ξ_u be the cluster of labels correspondent to the cluster of feature vectors $\mathbf{X}_u$:

$$\Xi_u = \{\xi_1, \dots, \xi_n, \dots, \xi_N\} \quad (7.3)$$

where each ξ_n is the phonetic label of feature vector $\mathbf{x}_n$, obtained through the use of a **FAM**. The number of different types of labels is Q , being $Q = 22$ in Spanish. In a general case:

$$\xi_n \in \{\iota_1, \dots, \iota_g, \dots, \iota_Q\} \quad (7.4)$$

Then, the **fGMM** is represented by $\Theta_{fGMM}^c = \{\mu_{fGMM,g}^c, \Sigma_{fGMM,g}^c\}_{g=1}^Q$, where $\mu_{fGMM,g}^c$ is the mean vector and $\Sigma_{fGMM,g}^c$ is the covariance matrix of Gaussian g in class c , although in this experimental set only the diagonal covariance matrices are used. To calculate the means and

the covariance matrix for each g , only the g^{th} frame phonetic group is used. For instance, to compute $\mu_{fGMM,1}^c$ and $\Sigma_{fGMM,1}^c$, only the feature vectors of class c labeled with ι_1 , (normally /a/), are used.

To compute the class membership of a certain feature vector $\mathbf{x}_n$ from the testing subgroup, contrasting the alternative hypothesis $H_{control}$ and H_{PD} , the likelihood ratio test is used:

$$\frac{p(\mathbf{x}_n | H_{PD})}{p(\mathbf{x}_n | H_{control})} \begin{cases} \geq \theta, & \text{accept } H_{PD} \\ < \theta, & \text{reject } H_{PD} \end{cases} \quad (7.5)$$

where

$$p(\mathbf{x}_n | H_c) = p_g^c(\mathbf{x}_n) \quad (7.6)$$

In this case, there is a single g for each $\mathbf{x}_n$, determined by the phonetic label associated to this feature vector. For instance, if the feature vector is labeled with /a/ ($g = 1$), $p(\mathbf{x}_n | H_c) = p_1^c(\mathbf{x}_n)$. It can be noticed that in this case, there is no Gaussian weightings as those used in the typical GMM models, as detailed in [195]. Since every single testing feature D -dimensional vector is unambiguously associated to a specific Gaussian density $p_g^c(\mathbf{x})$ (characterized by its mean vector $\mu_{fGMM,g}^c$ of length D and covariance matrix $\Sigma_{fGMM,g}^c$ of size $D \times D$), there is no need to use weightings to balance the use of the different $p_g^c(\mathbf{x})$.

For any feature vector $\mathbf{x}_n$ the Gaussian density $p_g^c(\mathbf{x}_n)$ is defined as:

$$p_g^c(\mathbf{x}_n) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}_n - \mu_{fGMM,g}^c)'(\Sigma_{fGMM,g}^c)^{-1}(\mathbf{x}_n - \mu_{fGMM,g}^c)\}}{(2\pi)^{D/2}|\Sigma_{fGMM,g}^c|^{1/2}} \quad (7.7)$$

Finally, the scores for each utterance $\mathbf{X}_u$ for the model of class c , containing N frames, are calculated by means of the log-likelihood of every frame as:

$$\Lambda_{fGMM,u}^c = \frac{1}{N} \log \prod_{n=1}^N p(\mathbf{x}_n | \Theta_{fGMM}^c) \quad (7.8)$$

being the global scores for each utterance:

$$\Lambda_{fGMM,u} = \Lambda_{fGMM,u}^{PD} - \Lambda_{fGMM,u}^{control} \quad (7.9)$$

Figure 7.4 contains a diagram of the training and testing processes of the fGMM scheme.

Figure 7.4: fGMM general scheme. Initially, the fGMM model is trained using the training data (training fold) and then, tested with the testing data (testing fold)

7.2.4 Gaussian Mixture Model - forced Universal Background Model (GMM-fUBM)

The Gaussian Mixture Model - forced Universal Background Model (GMM-fUBM) classification scheme is similar to the GMM-UBM but in this case the UBM is created using the phonetic labels from the UBM speech database as explained in section 7.2.3, obtaining a Forced Universal Background Model (fUBM). Afterwards, the final GMM-fUBM model is obtained performing the MAP adaptation of the UBM mean vectors by using the training utterances. In this case, the phonetic labels of the training database are not used and the adaptation is carried out in the traditional manner, as explained in [195].

Therefore, the idea is to obtain a UBM containing Gaussians directly associated to the various phonetic labels. This UBM will serve as the base for obtaining a new GMM model following traditional techniques.

Thus, after obtaining the fUBM model without weightings $\Theta_{fUBM} = \{\mu_g^{UBM}, \Sigma_g^{UBM}\}_{g=1}^Q$, the GMM-fUBM model will contain the UBM covariance matrix and the MAP adaptation of the UBM mean vectors, resulting in $\Theta_{fUBM}^c = \{\hat{\mu}_g^c, \Sigma_g^{UBM}\}_{g=1}^Q$.

To compute the class membership of a certain feature vector $\mathbf{x}_n$ from the testing subgroup, the procedure explained in section 6.2.3 is applied.

Figure 7.5 contains a diagram of the training and testing processes of the GMM-fUBM scheme.

7.2.5 Forced Gaussian Mixture Model - forced Universal Background Model (fGMM-fUBM)

The Forced Gaussian Mixture Model - forced Universal Background Model (fGMM-fUBM) is obtained after the adaptation mean vectors of the fUBM, but in this case the phonetic labels of the training database are also used. The idea behind this scheme is that each of the resulting mixture densities is obtained using exclusively segments of a certain

Figure 7.5: GMM-fUBM general scheme. Initially, the model is trained using the training data (training fold) and then, tested with the testing data (testing fold)

phonetic type (/a/, /b/, /d/, etc.), being able to compare more precisely these types between classes, as in the *fGMM* but using a *fUBM* which allows to generalize better.

Similarly to the previous section, after obtaining the *fUBM* model $\Theta_{fUBM} = \{\mu_g^{UBM}, \Sigma_g^{UBM}\}_{g=1}^Q$, the *fGMM-fUBM* model will be composed of the UBM covariance matrix and the adaptation of the UBM mean vectors using the training database and its correspondent phonetic labeling, resulting in $\Theta_{fUBM}^c = \{\hat{\mu}_{fGMM,g}^c, \Sigma_g^{UBM}\}_{g=1}^Q$.

This mean vectors adaptation is calculated as

$$\hat{\mu}_{fGMM,g}^c = \alpha_g^c \mu_{fGMM,g}^c + (1 - \alpha_g^c) \mu_g^{UBM} \quad (7.10)$$

where

$$\alpha_g^c = \frac{r_g^c}{r_g^c + \sigma} \quad (7.11)$$

being r_g^c the number of frames of class c in the training database labeled as ι_g and σ , is the relevance factor, as defined in [195].

Again, to compute the class membership of a certain feature vector $\mathbf{x}_n$ from the testing subgroup, the procedure explained in section 7.2.3 is applied.

Figure 7.6 contains a diagram of the training and testing processes of the *fGMM-fUBM* scheme.

Figure 7.6: fGMM-fUBM general scheme. Initially, the model is trained using the training data (training fold) and then, tested with the testing data (testing fold).

7.2.6 Optimized forced Gaussian Mixture Model - forced Universal Background Model (fGMM-fUBM)

Taking advantage of the phonetic labeling of the utterances and the unambiguous correspondence between the speech segments and the Gaussian mixtures, a modification of the previous scheme is proposed to optimize the models during the training. The idea behind this optimization is to recursively compute the accuracy of all the different types of training frames after obtaining the fGMM-fUBM classifier and use this accuracy as a weighting vector to change the importance of the frame types in the computation of class membership. In this sense, if a frame phonetic group in the training subset provides better accuracy than others, the weighting vector will make this type of frames more relevant in the final computation for a new testing utterance.

Therefore, after creating the fGMM-fUBM models, the membership of the training feature vectors to the two possible classes is computed and this information is used to create the weightings vector Φ^c :

$$\Phi^c \in \{\phi_1^c, \dots, \phi_g^c, \dots, \phi_Q^c\} \quad (7.12)$$

where:

$$\phi_g^c = \frac{n f_{rd,g}^c}{n f_{t,g}^c} \quad (7.13)$$

being $nf_{rd,g}^c$ the number of frames of phonetic group g and class c correctly classified, and $nf_{t,g}^c$ the total number of frames of phonetic group g and class c .

In this case, the new scores of any feature vector $\mathbf{x}_n$ respect to a certain model of class c will be

$$\lambda_{opt-fGMM-fUBM,n}^c = \log \left(p(\mathbf{x}_n | \Theta_{fGMM}^c) \right) \sum_{g=1}^Q \phi_g^c \eta_{g,n} \quad (7.14)$$

being $\eta_{g,n}$ an index indicating if the phonetic label associated to $\mathbf{x}_n$ is ι_g ($\eta_{g,n} = 1$) or not ($\eta_{g,n} = 0$). Therefore, the global scores for a certain utterance of N frames are:

$$\Lambda_{opt-fGMM-fUBM} = \frac{1}{N} \sum_{n=1}^N \left(\lambda_{opt-fGMM-fUBM,n}^{PD} - \lambda_{opt-fGMM-fUBM,n}^{control} \right) \quad (7.15)$$

These weights are used to evaluate the performance of the models and the obtained results yield new weights. The process is repeated until reaching a maximum number of iterations or until the overall accuracy does not increase more than a certain incremental improvement, ω_{min} .

7.3 METHODOLOGY

The purpose of the methodology of this experimental set is to analyze the four classification approaches exposed in section 7.2, for the automatic detection of PD from the speech. To carry out these tests, it is first necessary to perform the forced alignment of GITA, Neurovoz and Albayzin databases.

7.3.1 Forced Alignment

The FisherSP database is used to train a FAM using Kaldi. This database, sampled at 8 kHz is normalized and divided into frames of 25 ms with a frame shift of 10 ms. A vector containing 20 MFCC, plus $\Delta + \Delta\Delta$ acoustic features is processed from each frame. To do the initial monophone training-alignment, a subset of 10.000 utterances is randomly selected, producing the monophone model (40 iterations). After this, an alignment is performed in a subset of 30.000 utterances to obtain new segments for the the second round of training-alignment. Afterwards, two rounds of 30 iterations are performed to train a triphone model using 30.000 utterances followed by the correspondent alignments. Then, and now employing 100.000 utterances, a round of 35 iterations employing a LDA+MLLT transformation of the acoustic features as input

is performed. Finally, 30 iterations of the training-aligning sequence using **fMLLR** transformation of the features produce the final **FAM**.

This **FAM** is used to segment and label all the 6 fixed sentences of the GITA database, 6 fixed sentences on Neurovoz database and all the recordings of the corpus 1 of Albayzin (4.801 fixed sentences), employing in all the cases, their correspondent transcriptions as metadata. To do this, all of these databases are normalized, filtered and downsampled to 8 kHz, as the database used to train the **FAM** and characterized with 20 **MFCC**+ Δ + $\Delta\Delta$ features per frame (25 ms, 10 ms frame shift).

7.3.2 Forced Gaussian Mixture models

A first round of trials was carried out to train and test **fGMM** models with GITA and Neurovoz databases separately. In the second round, **GMM-fUBM** models were trained and tested separately for GITA, Neurovoz and CzechPD, using Albayzin for the **UBM**. In the third and fourth rounds, **fGMM-fUBM** and optimized **fGMM-fUBM** models respectively, were trained and tested separately using, in this case GITA and Neurovoz databases, employing Albayzin to create the **UBM**. For the optimized **fGMM-fUBM**, the maximum number of iterations was limited to 20 and $\omega_{min} = 0.5\%$. In all the cases, the recordings were parameterized using Rasta-**PLP** + Δ + $\Delta\Delta$ with a number of **PLP** coefficients, F , varying in the range $\{10...20\}$ with steps of 2. The used frame length was 15 ms and the number of coefficients in the **FIR** filter, 5. Respecting the speech tasks, all the available fixed text tasks are employed jointly to train the respective GITA and Neurovoz models in all the trials. However, in the second round of trials, the **DDK** task and monologues were utilized too, while only the **DDK** task from CzechPD was employed. This last database was used in this round since the **GMM-fUBM** scheme does not require forced alignment on the training database (forced alignment was only performed in fixed texts from databases in Spanish language). The validation of the models was carried out using a k-folds cross-validation scheme, with $k=11$. Therefore, for each round of trials 11 models were trained and tested per database, speech task and F . It is important to remark that the Rasta-**PLP** + Δ + $\Delta\Delta$ coefficients are calculated for the whole utterance and, after obtaining the feature vectors, these are distributed into the correspondent phonetic bins for training or testing, attending to their forced alignment labeling.

Finally, as in the previous experimental set, a round of cross-database trials is performed using the **GMM-fUBM** scheme with **DDK** tasks from GITA, Neurovoz and CzechPD. These trials consist in the combination of two of the parkinsonian databases to train a model which is tested with the remaining database. This process is repeated for all the possible combinations.

7.4 RESULTS

In this section, the results from the four rounds of trials are detailed. Table 7.1 shows the results of the analysis of the fGMM scheme, Table 7.2 shows the results related to the GMM-fUBM models while Tables 7.3 and 7.4 include the results referred to the fGMM-fUBM and optimized fGMM-fUBM classification schemes, respectively. All of these tables include the classification accuracy $\pm$ CI, AUC, sensitivity, specificity, N , and the Spearman’s correlation of the scores with the UPDRS and H-Y scales with their respective p-values.

Best results for each database in the k-folds cross-validation are in bold.

Table 7.1: fGMM results

Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed text	79 ± 8	0.88	0.76	0.82	10	0.34	0.0211	0.2	0.176
Neurovoz	Fixed text	71 ± 10	0.86	0.67	0.77	14	0.5	0.0021	0.34	0.0346

Table 7.2: GMM-fUBM results

Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed text	78 ± 8	0.88	0.78	0.78	12	0.06	0.6837	0.05	0.7314
	DDK	79 ± 8	0.86	0.86	0.72	12	0	0.9797	0	0.9955
	Monol.	78 ± 8	0.84	0.73	0.82	10	0.38	0.0065	0.25	0.0819
Neurovoz	Fixed text	81 ± 9	0.87	0.83	0.78	14	0.3	0.0802	0	0.988
	DDK	81 ± 9	0.85	0.83	0.77	20	0.29	0.0937	0.14	0.3992
	Monol.	66 ± 14	0.67	0.35	0.83	10	0.17	0.5732	0.06	0.8513
CzechPD	DDK	94 ± 1	0.97	0.9	1	20	0.04	0.8534	0.01	0.9553

Table 7.3: fGMM-fUBM results

Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed text	81 ± 8	0.88	0.84	0.78	10	0.32	0.0318	0.21	0.1528
Neurovoz	Fixed text	75 ± 10	0.86	0.76	0.74	14	0.42	0.011	0.34	0.0331

Table 7.4: Optimized fGMM-fUBM results

Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed text	81 $\pm$ 8	0.88	0.84	0.78	10	0.32	0.0307	0.22	0.1504
Neurovoz	Fixed text	79 $\pm$ 9	0.87	0.8	0.77	12	0.42	0.012	0.36	0.0225

Figure 7.7: ROCCH-DET curves obtained employing testing scores from GITA in the classification schemes: fGMM, GMM-fUBM, fGMM-fUBM and optimized fGMM-fUBM. Fixed text is used as speech task.

Figures 7.7 and 7.8 plot the ROCCH-DET curves for the GITA and Neurovoz databases respectively, using the four proposed classification schemes.

Figures 7.9 and 7.10 show the sensitivity and specificity per frame phonetic group in GITA and Neurovoz databases respectively, employing fixed text task in an optimized fGMM-fUBM scheme. In these figures, the symbols B and D are fricatives while b and d are plosives. R is multiple vibrant and r is simple vibrant.

Table 7.5 includes the results of the cross-database tests using GITA, Neurovoz and CzechPD as testing databases separately. For each case, the other two parkinsonian databases are used to train the model.

Finally, Figure 7.11 plots the ROCCH-DET curves of the three parkinsonian databases using the DDK task in a GMM-fUBM classification scheme.

7.5 DISCUSSION

In this experimental set, several approaches are proposed to create a GMM in such a manner that, after obtaining the final model, each Gaussian mixture models only a phonetic segment. To obtain the gaussian densities within the GMM, the available phonetic segments of the

Figure 7.8: ROCCH-DET curves obtained employing testing scores from Neurovoz in the classification schemes: fGMM, GMM-fUBM, fGMM-fUBM and optimized fGMM-fUBM. Fixed text is used as speech task.

speech are employed, obtained with speech forced alignment techniques.

All the approaches are based on the use of GMM-UBM techniques but with differences in the material used to train the UBM and the final models. These approaches proposed here are: fGMM, GMM-fUBM, fGMM-fUBM and optimized-fGMM-fUBM.

For GITA, the optimized fGMM-fUBM approach provides the best results (accuracy equals to 81% and AUC equals to 0.88) while for Neurovoz and CzechPD, the GMM-fUBM scheme is the one yielding the highest accuracies and AUC (accuracy equals to 81% and 94% and AUC equals to 0.87 and 0.97, respectively). In this case, as in the previous experimental set, the CI intervals for the best results of GITA and Neurovoz are usually over $\pm 9\%$ which produces overlapping among the final result ranges. Therefore, the conclusions that can be inferred by analyzing the results after employing these databases are focused on the observed trends more than in absolute terms.

It can be considered that the scheme GMM-fUBM provides, in general lines, the best results since, for GITA and Neurovoz produces the best AUC and similar accuracies. In addition, this last scheme does not require to use forced alignment in the parkinsonian database, allowing the possibility to create supervised or unsupervised schemes, depending on the speech material used (fixed text and DDK will produce supervised approaches while monologues, unsupervised approaches).

(a) Sensitivity and specificity per frame phonetic group in GITA using training scores.

(b) Sensitivity and specificity per frame phonetic group in GITA using testing scores.

Figure 7.9: Per frame sensitivity and specificity in GITA database using fixed text in an optimized fGMM-fUBM scheme.

Table 7.5: Cross-database GMM-fUBM results using DDK task. For each test database, the remaining two parkinsonian databases are used to train the model.

Test Database	Speech task	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	DDK	66 ± 9	0.76	0.9	0.42	18	0.22	0.1316	0.11	0.4767
Neurovoz	DDK	74 ± 10	0.78	0.87	0.5	10	0.2	0.2557	0.09	0.5822
CzechPD	DDK	76 ± 14	0.87	0.95	0.5	16	0.12	0.6066	0.07	0.7792

The cross-database results found in Table 7.5 demonstrate a reduced accuracy compared with the k-folds cross validations shown in Tables 7.1, 7.2, 7.3 and 7.4. However, the sensibility in the detection of PD is maintained in all the cases.

(a) Sensitivity and specificity per frame phonetic group in Neurovoz using training scores.

(b) Sensitivity and specificity per frame phonetic group in Neurovoz using testing scores.

Figure 7.10: Per frame sensitivity and specificity in Neurovoz database using fixed text in an optimized fGMM-fUBM scheme.

Figure 7.11: ROCCH-DET curves obtained employing testing scores GITA, Neurovoz and CzechPD in the GMM-fUBM classification scheme. DDK sequence (/pa-ta-ka/) is used as speech task.

Observing Figures 7.9 and 7.10 it is possible to deduce that the phonetic unities leading to better accuracies using GITA are /D/, /g/, /h/ (aspirated h), /R/, /w/ and /y/ while /B/, /d/, /k/, /m/ and /r/ provide the lowest accuracies. Regarding Neurovoz, /C/, /D/, /i/, /l/, /p/ and /w/ produce the highest accuracies while /a/, /j/ and /m/, the lowest. In both databases, phonetic units requiring a higher narrowing of the vocal tract but without a burst, such as the units /C/, /D/, /g/ and /R/ tend to be more decisive in the detection while others such as /B/ or /m/ are less influential. Regarding vowels, /u/ and /w/ produce good results in both databases. The differences between the results on the two databases is caused by the dialectal differences between Castillian and Colombian Spanish and by the position of the phonetic unities within the sentences, which influences its articulation. In general terms, these figures suggest that the differences in accuracy between the distinct specific phonetic units is variable and there is not a phonetic unit or groups of units that can clearly be used separately to detect PD. This suggest that PD may affect to the articulatory sequence as a whole with a higher influence on phonetic units requiring a narrowing of the vocal tract but without a burst and with a lower influence in nasals. However, more general categories of articulatory movements related with the type of narrowing of the vocal tract or the use of the glottal source must be performed.

7.6 SUMMARY

In this experimental set several approaches, namely fGMM, GMM-fUBM, fGMM-fUBM and optimized-fGMM-fUBM are proposed for the automatic detection of PD. The obtained models contain Gaussian densities created or adapted employing only specific phonetic units, allowing to compare precisely their features between patients and controls.

Results suggest that PD affects to the articulatory sequence as a whole, influencing more clearly phonetic units requiring a higher narrowing of the vocal tract.

Cross-validation trials (k-folds) provide accuracies between 81% and 94%, with AUC between 0.87 and 0.97 depending on the database, while cross-database trials provide accuracies between 66% and 76% with AUC between 0.76 and 0.87.

EXPERIMENTAL SET 3: ALLOPHONIC DISTILLATION

8.1 INTRODUCTION

The study of literature reveals some prospective works, such as [48], in which authors analyze perceptually the dysfunctions of 200 untreated patients, showing that PD affects in a different manner to the perceived groups of allophones, being stop-plosives, fricatives and affricates the most affected. To this respect, Figures 8.1 and 8.2 allow to establish a comparison between the waveforms and spectrograms of two controls and two parkinsonian speakers while pronouncing the same word, containing three plosives (/pe-ta-ka/). In Figure 8.1, a newly diagnosed patient is compared with an age-matched control. It is possible to observe that, although the patient's waveform exhibits voice bars between the end of the vowels and the beginning of the plosives, the spectrogram shows a tendency to convert the burst of plosives to a more gradual articulation, like in fricatives. This effect is not observed in the control speaker, where the bursts of plosives are clearly marked, specially in the spectrogram. More pronounced effects can be found in Figure 8.2 where the voice bars are shorter in the patient (before /t/) or nonexistent (before /k/), transforming the plosive consonant /k/ into something similar to a fricative /g/. The waveform segment substituting the voice bar reveals a possible lack of control of the glottal source which keeps vibrating in an articulatory moment where it should be stopped. These effects are visible in the spectrogram too, where the bursts of the plosives are almost indistinguishable and, in the case of the plosive /t/, the first frequency peak from the previous vowel /a/ is joined with the first formant of the following vowel /a/. In examples like this, a statistical modeling of the acoustic features of plosive segments in patients and controls, separately, will lead to substantially different probabilistic densities.

Thus, so as to analyze the use of different types of segments of speech in an automatic detector and to confirm the influence of PD in the different types of segments, in this experimental set several PD detection approaches were analyzed, making use of the the GMM-UBM techniques and the PLP features described in the experimental set 1 but using only certain segments of speech. These segments were selected depending on the manner of articulation in a process coined in this thesis: *allophonic distillation*. For these purposes, state-of-the-art forced alignment techniques are used.

This type of experimental set at segmental level is aimed to study the muscle coordination taking part in the speech production of patients.

(a) Idiopathic PD female speaker. Age: 59. UPDRS: 9. Span: 720 ms

(b) Control female speaker. Age: 59. Span: 857 ms

Figure 8.1: Waveforms and spectrograms of a parkinsonian (newly diagnosed) and a control speaker pronouncing the word "petaca". Obtained from the Neurovoz database.

8.2 THEORETICAL BACKGROUND

8.2.1 *Allophonic distillation*

Different categorizations of allophones of the Spanish language can be found in literature, from which the ones proposed by [222] are widely used. These categorizations can be found in table 8.1.

To this respect, the plosive consonants are those preceded by a stop or a total obstruction of the articulators produce a burst after its release. An example of a plosive is the allophone p from the phoneme $/p/$ in the word "pastel". Fricatives are these sounds in which the constriction is not complete and there is not a stop. An example of fricative is the allophone f , from the phoneme $/f/$ in the word "alféizar". Affricates are these sounds which are preceded by a stop, have a burst and end with a constriction, as the fricatives. An example is the allophone $'tʃ$ in the word "colchón". Liquids are similar to these last but in this case the articulators do not approach enough to produce the same turbulence as in fricatives. The allophone $'r$ from the phoneme $/r/$ in the word "rey" is an example of liquid. Nasals are produced when the soft palate is lowered and allows the air coming from the larynx to pass through the nasal cavities and escape through the nose. As a difference with the previous types of consonants in Spanish, nasals

(a) Idiopathic PD female speaker. Age: 85. UPDRS: 47. Span: 810 ms

(b) Control female speaker. Age: 83. Span: 780 ms

Figure 8.2: Waveforms and spectrograms of a parkinsonian (in an advanced stage) and a control speaker pronouncing the word "petaca". Obtained from Neurovoz database.

are sonorants, which means that while their articulation, the glottal source is functioning. In Spanish, the phonemes /p/, /t/ and /k/ are always plosives while /b/, /d/ and /g/ can be plosives (allophones b, d, g) or fricatives (allophones β , δ , γ) depending on the precedent phonemes. For instance, the phoneme /b/ is plosive when used in the word "barco" pronounced alone or at the beginning of a sentence after a silence, but becomes fricative after the phoneme /n/ in the word "cambiar". In the first case, there is a stop and a burst and, in the second case, the closure of the articulators is not total and there is no stop, resulting in a fricative.

Figure 8.3 shows an example of a forced alignment of the sentence "La petaca blanca" including the different types of allophones obtained after the alignment.

The categorization referent to the manner of articulation (3 in Table 8.1) is going to be used in this experimental set. Thus, the use of different segments included in the consonant categories will allow to assess the influence of PD on different ranges of constrictions during articulation (plosives, fricatives and liquids), the use of the soft palate or the hyponasalization or hypernasalization¹ effects (nasals) or the vi-

¹ The hyponasalization effect appears when there is not nasalization during articulation but there should be, while hypernasalization is the opposed effect [231].

Table 8.1: Categorizations of allophones suggested in [222]

1. Depending on the action of vocal folds	a. Sonorant sounds	
	b. Obstruent or voiceless sounds	
2. Depending on the action of the soft palate	a. Oral sounds	
	b. Nasal sounds	
3. Depending on the manner of articulation	a. Vowels	
	b. Consonants:	i. Plossives
		ii. Fricatives
		iii. Affricates
		iv. Nasals
	v. Liquids	
4. Depending on the point of articulation	a. Vowels	
	b. Consonants:	i. Labials
		ii. Labiodentals
		iii. Dentals
		iv. Interdentals
		v. Alveolars
		vi. Palatals
		vii. Velars

bration of the vocal folds in combination with articulatory movements (vowels).

In the same manner as in the previous experimental set, it is possible to use **FAM** to label **specific allophonic segments** in the speech recordings. Then, these labels can be employed to select the speech segments coming from a certain manner of articulation, it is to say, segments containing only affricates, fricatives, liquids, nasals, plosives or vowels. This process of selection of groups of phonemes has been coined in this thesis as *allophonic distillation*.

Hence, if an utterance is composed of N frames, the cluster of feature vectors, $\vec{X}_u$ is defined as:

$$\vec{X}_u = \{\vec{x}_1, \dots, \vec{x}_n, \dots, \vec{x}_N\} \quad (8.1)$$

after a distilling, the new cluster of feature vectors, $\vec{X}_{u-dist}$, will be:

$$\vec{X}_{u-dist} = \{\vec{x}'_1, \dots, \vec{x}'_n, \dots, \vec{x}_A\} \quad (8.2)$$

where $A \leq N$, being A , the total number of frames selected after allophonic distillation.

Figure 8.3: Example of forced alignment labeling obtained with Kaldi, including waveform, spectrogram, phonemes, phoneme category and transcription. LIQ stands for liquid; VOW, for vowel; PLOS, for plosive, FRIC, for fricative; and NAS, for nasal.

8.3 METHODOLOGY

The methodology of this experimental set includes several parts which are detailed in this section. As an introduction, it can be said that firstly, a forced alignment model in Spanish language was trained. This model was used to segment and label three databases: GITA, Neurovoz and Albayzin. With the first two, several [GMM-UBM](#) models were trained and analyzed, employing the third one, Albayzin, as [UBM](#). These models were obtained applying, in turns, allophonic distillation in the parkinsonian or auxiliary databases employed to train the [UBM](#) (or in both) in order to analyze the importance of the different types of segments for the automatic detection of [PD](#). When possible, some trials using CzechPD were performed too. Finally, fusion of scores and cross-database trials complete and validate, respectively, the proposed methodology.

Regarding the creation of the Spanish [FAM](#) and the segmentation and phonetic labeling of GITA, Neurovoz and Albayzin, both tasks were accomplished as explained in the previous experimental set (section [7.3.1](#)).

Once the forced alignment is made, several types of trials were carried out using only specific allophonic segments from the speech to train [GMM-UBM](#) models, after the allophonic distillation. Three different approaches were employed, depending on where the allophonic distillation was applied: in the adaptation-testing (parkinsonian speech) or in the [UBM](#) databases. On each trial, a model was trained and tested following a k-folds cross-validation strategy (11 folds) using only one type of specific allophonic distillation: fricatives, liquids, nasals, plosives and vowels. Affricate segments were not analyzed as they are clearly underrepresented in GITA and Neurovoz fixed texts, as it can be inferred from Table [8.2](#). In all the cases, the [UBM](#) was trained and adapted as explained in section [6.2.3](#).

Table 8.2: Total number of repetitions of consonants in GITA and Neurovoz fixed texts.

Type of consonant	# repetitions	
	GITA	Neurovoz
Affricate	1	1
Fricative	30	35
Liquid	23	26
Nasal	19	21
Plosive	26	35

In the first approach, distillation was applied in the adaptation-testing database. On the trials associated to the analysis of this approach, GITA and Neurovoz were employed separately. From these databases, only fixed texts tasks were used. A depiction of this first approach is shown in Figure 8.4.

Figure 8.4: First approach followed for the allophonic distillation methodology.

For the second approach, a new round of trials was repeated following the same premises but pursuing now allophonic distillation in the UBM too. A depiction of the second approach is presented in Figure 8.5.

Figure 8.5: Second approach followed for the allophonic distillation methodology.

It is convenient to remark that, in these two first approaches, the distillation is applied to both, training and testing utterances in the parkinsonian databases. In this way, the created systems modeled only specific phonetic segments (fricative, plosive, etc.) and during the testing stage, only their respective specific segments were employed too. Therefore, the two first approaches and their associated rounds of trials allow to analyze the importance of the different types of phonetic segments in the automatic detection of PD using running speech.

In the third approach, allophonic distillation was applied only in the UBM, in order to analyze the importance of the initialization of the GMM-UBM model, as the distillation in the UBM database produces GMM-UBM models more oriented to fricatives, liquids, nasals, plosives or vowels, as appropriate. In this last case, all three parkinsonian databases were employed. In this case, for GITA and Neurovoz, three different acoustic materials were examined separately: fixed texts, monologues, and DDK tasks. As no Czech FAM has been trained in this thesis to distill a language specific UBM database, only Albayzin was used and distilled. Therefore, exclusively the DDK task from CzechPD was considered. Figure 8.6 portrays a diagram of these types of trials.

Figure 8.6: Third approach followed for the allophonic distillation methodology.

In all the three approaches, the fixed texts were not studied separately, sentence by sentence, but all together to train the same model within the trials associated to one specific database. In all the cases, the same front-end of the previous experimental set was used where all the utterances are normalized, and characterized using Rasta-PLP, with F varying in the range $\{10...20\}$ in steps of 2. Regarding the classification, G varied in powers of 2 from 4 to 256. In this experimental set, as in the rest of experimental sets described in this thesis, none of the utterances or even frames from a speaker employed to train the models was used in the testing stage during the cross-validation. In the same manner, the Rasta-PLP + $\Delta + \Delta\Delta$ coefficients are calculated for the whole utterances and, after obtaining the feature vectors, these are distributed into the correspondent allophonic categories for training or testing, attending to their forced alignment labeling.

After analyzing the results of all approaches, a fusion of scores from the approach providing the best accuracy and AUC was studied. To obtain a final score coming from the fusion of several scores, a logistic regression was employed. For each database, the speaker scores obtained from the five possible distillations (fricative, liquid, nasal, plosive and vowels) were fused following all the possible combinations

of n – tuples, going from 2 – tuples to 5 – tuples. Therefore, for each combination relative to a certain speaker, a new score was calculated considering only the scores coming from this speaker and having a common F and G . For instance, in the combination fricative-liquid-vowels of a certain trial employing 12 Rasta-PLP coefficients and 256 Gaussians, the three single scores for each speaker coming from the fricative, liquid and vowel distillations, respectively, were used for the fusion.

Finally, a last round of trials followed a cross-database validation in the third approach, in which the Albayzin database was distilled again to obtain the five different types of UBM, and the DDK tasks from the parkinsonian databases were used to train and test the models. Particularly, three laps of trials were carried out. On each one, two of the databases were used to train the model and the remaining was used exclusively in the testing process. Therefore, there is a model created with all the speakers contained in GITA and Neurovoz, and tested with CzechPD; other model created with GITA and CzechPD, and tested with Neurovoz; and a third model trained employing the utterances from Neurovoz and CzechPD, and tested with GITA. In this case, the same front-end and classification parameters of the rest of the experimental set were used.

8.4 RESULTS

Table 8.3 includes the best results of the first approach, where allophonic distillation is applied only to the parkinsonian databases while Table 8.4 reflect the best results in the second approach, where the distillation is applied to the parkinsonian and to the UBM database (Albayzin).

Table 8.3: Approach 1: results after applying allophonic distillation to the parkinsonian database and no distillation to the Albayzin database (UBM). Only fixed text is used as speech task.

Database	Distill.	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fricatives	77 $\pm$ 8	0.85	0.72	0.82	10	128	0.38	0.0119	0.26	0.0931
	Liquids	77 $\pm$ 8	0.84	0.77	0.78	14	16	0.06	0.7019	0	0.9844
	Nasals	77 $\pm$ 8	0.81	0.7	0.84	12	64	0.13	0.3932	0.15	0.3376
	Plosives	84 $\pm$ 7	0.9	0.83	0.84	16	128	0.07	0.6414	0.12	0.4477
	Vowels	81 $\pm$ 8	0.89	0.74	0.88	12	128	0.22	0.1525	0.12	0.4274
Neurovoz	Fricatives	83 $\pm$ 9	0.88	0.86	0.78	10	256	0.3	0.0905	0.15	0.3872
	Liquids	82 $\pm$ 9	0.91	0.81	0.83	10	64	0.55	0.0011	0.41	0.0143
	Nasals	82 $\pm$ 9	0.89	0.83	0.78	12	8	−0.04	0.8198	−0.15	0.3989
	Plosives	83 $\pm$ 9	0.94	0.81	0.87	14	64	0.38	0.031	0.1	0.5841
	Vowels	77 $\pm$ 10	0.89	0.71	0.87	14	256	0.51	0.0028	0.39	0.0209

Tables 8.5, 8.6 and 8.7 include the best results for GITA, Neurovoz and CzechPD, respectively, in the third approach. Best results per database, speech tasks and approach are shaded in green and best global results per database are in bold. Figures 9.11 and 9.12 include a graphical representation of best accuracies and AUC reached in the different trials. Finally, Figure 8.8 include the ROCCH-DET curves of the trials with best accuracy of the three different approaches for GITA and Neurovoz, employing fixed texts. Figure 8.8 shows the ROCCH-DET curves of the different types of segments in the third approach with the three parkinsonian databases and Figure 8.9 shows the curves in the cross-database trials.

Table 8.4: Approach 2: results after applying allophonic distillation to the parkinsonian and Albayzin databases (UBM). Only fixed text is used as speech task.

Database	Distill.	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fricatives	79 $\pm$ 8	0.86	0.74	0.84	14	32	0.28	0.0642	0.22	0.1601
	Liquids	77 $\pm$ 8	0.83	0.79	0.76	18	4	0.06	0.6932	-0.1	0.5169
	Nasals	77 $\pm$ 8	0.84	0.79	0.76	12	16	0.13	0.3907	-0.02	0.8751
	Plosives	85 $\pm$ 7	0.89	0.81	0.88	18	128	0.06	0.6757	0.16	0.3093
	Vowels	86 $\pm$ 7	0.9	0.79	0.92	14	128	0.22	0.1533	0.13	0.3854
Neurovoz	Fricatives	86 $\pm$ 8	0.87	0.88	0.83	20	16	0.28	0.1228	0.06	0.7329
	Liquids	83 $\pm$ 9	0.9	0.81	0.87	16	32	0.57	0.0006	0.49	0.0028
	Nasals	82 $\pm$ 9	0.9	0.83	0.78	16	16	0.29	0.1128	0.09	0.5993
	Plosives	85 $\pm$ 9	0.93	0.81	0.91	12	256	0.34	0.0598	0.1	0.5687
	Vowels	77 $\pm$ 10	0.89	0.76	0.78	10	128	0.52	0.0025	0.43	0.0107

(a) ROCCH-DET curves in GITA.

(b) ROCCH-DET curves in Neurovoz.

Figure 8.7: ROCCH-DET curves for GITA and Neurovoz including the scores of the best results of each approach for these databases. In all the cases, fixed text is acoustic material, as these produce the best results.

Table 8.5: Approach 3: results after applying allophonic distillation to the Al-bayzin database (UBM). GITA (with no allophonic distillation) is used to train and test the models.

Speech task	Distill.	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
Fixed Text	Fricatives	82 ± 8	0.89	0.82	0.82	10	128	0.09	0.5518	0.02	0.8837
	Liquids	81 ± 8	0.88	0.74	0.88	12	128	0.14	0.3398	0.06	0.6838
	Nasals	82 ± 8	0.88	0.82	0.82	12	128	0.13	0.3898	0.08	0.5741
	Plosives	85 ± 7	0.91	0.82	0.88	12	128	0.13	0.4016	0.08	0.6049
	Vowels	82 ± 8	0.89	0.76	0.88	10	256	0.13	0.3931	0.03	0.8502
DDK	Fricatives	80 ± 8	0.87	0.8	0.8	16	32	0.01	0.9455	-0.05	0.7504
	Liquids	82 ± 8	0.87	0.8	0.84	14	256	-0.03	0.8505	-0.03	0.8322
	Nasals	83 ± 7	0.89	0.86	0.8	10	32	-0.04	0.8018	-0.07	0.6511
	Plosives	82 ± 8	0.88	0.86	0.78	10	8	0.01	0.9698	-0.02	0.9065
	Vowels	83 ± 7	0.88	0.86	0.8	14	16	0.15	0.3226	0.05	0.7526
Monol.	Fricatives	80 ± 8	0.87	0.71	0.88	12	128	0.36	0.0114	0.25	0.0798
	Liquids	80 ± 8	0.87	0.73	0.86	12	64	0.43	0.0021	0.31	0.0324
	Nasals	77 ± 8	0.84	0.69	0.84	20	256	0.35	0.0148	0.26	0.073
	Plosives	80 ± 8	0.88	0.71	0.88	10	128	0.4	0.0039	0.26	0.0674
	Vowels	78 ± 8	0.84	0.76	0.8	14	128	0.37	0.0094	0.23	0.1099

Table 8.6: Approach 3: results after applying allophonic distillation to the Al-bayzin database (UBM). Neurovoz (with no allophonic distillation) is used to train and test the models.

Speech task	Distill.	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
Fixed Text	Fricatives	89 ± 7	0.93	0.87	0.91	12	256	0.4	0.0159	0.19	0.2381
	Liquids	87 ± 7	0.93	0.87	0.88	14	128	0.43	0.0107	0.14	0.3889
	Nasals	85 ± 8	0.93	0.85	0.84	14	128	0.47	0.0042	0.19	0.2556
	Plosives	86 ± 8	0.92	0.85	0.88	18	256	0.44	0.0078	0.18	0.2694
	Vowels	85 ± 8	0.92	0.83	0.88	12	128	0.37	0.0271	0.11	0.5158
DDK	Fricatives	83 ± 9	0.9	0.89	0.73	10	64	0.18	0.317	0.15	0.3533
	Liquids	81 ± 9	0.89	0.85	0.73	10	32	0.25	0.1605	0.2	0.2311
	Nasals	82 ± 9	0.87	0.85	0.77	18	64	0.22	0.2063	0.19	0.2624
	Plosives	86 ± 8	0.88	0.89	0.81	14	64	0.19	0.2763	0.07	0.6861
	Vowels	81 ± 9	0.88	0.87	0.69	14	64	0.24	0.1701	0.13	0.4504
Monol.	Fricatives	74 ± 12	0.77	0.47	0.9	16	64	0.52	0.0699	0.47	0.1078
	Liquids	55 ± 14	0.6	0.53	0.57	20	4	0.34	0.2609	0.43	0.14
	Nasals	70 ± 13	0.65	0.41	0.87	12	64	0.16	0.5927	-0.06	0.8467
	Plosives	70 ± 13	0.73	0.41	0.87	12	32	0.39	0.1895	0.27	0.3689
	Vowels	72 ± 13	0.74	0.65	0.77	16	32	0.52	0.071	0.42	0.1488

Table 8.7: Approach 3: results after applying allophonic distillation to the Al-bayzin database (UBM). CzechPD (with no allophonic distillation) is used to train and test the models. Only DDK task is considered.

Speech task	Distill.	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
DDK	Fricatives	94 $\pm$ 1	0.96	0.9	1	18	64	-0.06	0.7992	-0.04	0.8659
	Liquids	94 $\pm$ 1	0.95	0.9	1	20	64	0.02	0.9401	0.01	0.9796
	Nasals	91 $\pm$ 1	0.95	0.9	0.93	16	64	0.03	0.8911	0.01	0.9611
	Plosives	91 $\pm$ 1	0.98	0.9	0.93	16	16	0.04	0.8689	0.07	0.7712
	Vowels	91 $\pm$ 1	0.96	0.9	0.93	16	32	0.02	0.9383	-0.05	0.8419

Table 8.8: Best results after the fusion of scores for the three parkinsonian databases separately. The scores were obtained after applying the third approach (allophonic distillation only in the UBM database)

Database	Speech task	Combination	Accuracy $\pm$ CI	AUC	Sens.	Spec.	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed text	plosive-liquid	84 $\pm$ 7	0.9	0.8	0.88	0.12	0.4058	0.12	0.4323
	DDK	nasal-liquid	83 $\pm$ 7	0.88	0.84	0.82	0.05	0.7589	0.01	0.9587
	Monol.	liquid-vowel	82 $\pm$ 8	0.89	0.8	0.84	0.42	0.0025	0.28	0.0549
Neurovoz	Fixed text	fricative-vowel	89 $\pm$ 7	0.95	0.89	0.88	0.34	0.0435	0.24	0.1368
	DDK	liquid-vowel	83 $\pm$ 9	0.89	0.87	0.77	0.16	0.367	0.15	0.3537
	Monol.	plosive-nasal-vowel	77 $\pm$ 12	0.79	0.53	0.9	0.45	0.1234	0.2	0.5114
CzechPD	DDK	fricative-nasal	94 $\pm$ 1	0.98	0.9	1	-0.06	0.8044	-0.19	0.4152

Figure 8.8: ROCCH-DET curves for the three parkinsonian speech databases in the third approach (allophonic distillation applied only on UBM database, Albayzin). Fixed text are used for GITA and Neurovoz database, and DDK for CzechPD.

8.5 DISCUSSION

In this experimental set, three different approaches based on [GMM-UBM](#) classification schemes were tested. In each scheme, an allophonic distillation (fricative, liquid, nasal, plosive and vowel) was applied to the parkinsonian database or to the one employed for the [UBM](#). In general, as in previous experimental sets, the [CI](#) of the accuracy values generates overlapping margins in the results, which implies that the analysis of these can be made in terms of trends but does not allow to obtain definitive or clearly circumscribed conclusions.

8.5.1 Allophonic distillations

Regarding the first approach, [Table 8.3](#) shows that best results (accuracy, [AUC](#) and sensitivity) are obtained for plosive and vowel distillation in GITA and fricative and plosive distillation in Neurovoz. In the

Table 8.9: Cross-database results in GITA, Neurovoz and CzechPD, employing Albayzin for the UBM with the five different types of allophonic distillation. For every trial, two parkinsonian databases are used for training and the remaining, for testing.

Test Database	Distill.	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
Gita	Fricatives	74 $\pm$ 9	0.86	0.66	0.82	14	32	0.12	0.4408	-0.09	0.5361
	Liquids	69 $\pm$ 9	0.79	0.94	0.44	18	256	0.26	0.0728	0	0.9923
	Nasals	75 $\pm$ 8	0.84	0.86	0.64	20	32	-0.05	0.729	-0.13	0.3762
	Plosives	74 $\pm$ 9	0.8	0.86	0.62	16	32	0.1	0.4879	-0.07	0.6369
	Vowels	72 $\pm$ 9	0.81	0.7	0.74	20	4	0.04	0.7849	0.08	0.5804
Neurovoz	Fricatives	75 $\pm$ 10	0.85	0.76	0.73	20	32	0.36	0.0354	0.23	0.1686
	Liquids	74 $\pm$ 10	0.79	0.87	0.5	14	4	0.11	0.53	0.09	0.5862
	Nasals	74 $\pm$ 10	0.82	0.78	0.65	20	256	0.22	0.2121	0.13	0.4533
	Plosives	72 $\pm$ 10	0.78	0.93	0.35	20	128	0.37	0.0333	0.35	0.0308
	Vowels	81 $\pm$ 9	0.83	0.91	0.62	18	32	0.32	0.0616	0.22	0.1891
CzechPD	Fricatives	79 $\pm$ 14	0.93	0.9	0.64	10	16	0.13	0.5894	0.17	0.486
	Liquids	82 $\pm$ 13	0.9	0.9	0.71	10	8	0.34	0.1431	0.21	0.3764
	Nasals	82 $\pm$ 13	0.86	0.95	0.64	12	4	0.37	0.1035	0.09	0.7004
	Plosives	79 $\pm$ 14	0.87	0.95	0.57	12	16	0.08	0.7282	-0.02	0.9285
	Vowels	82 $\pm$ 13	0.95	0.85	0.79	10	4	0.21	0.3756	-0.03	0.9163

second approach, where the distillation is applied to both, parkinsonian and UBM databases, best results are obtained, again, for vowel and plosive distillation in the case of GITA and for fricative and plosive, in the case of Neurovoz. In the case of GITA, the differences between these distillations (plosives and vowels) and the other three are more accused than in the case of Neurovoz (plosives and fricatives).

In these two approaches occurs that, although vowel segments provide good results in GITA, this behavior is not clearly identified in Neurovoz. Similarly, although fricatives provide reasonably good results in the Neurovoz trials, lower accuracy and AUC is found in the GITA trials. However, plosive segments have a similar accuracy, AUC and sensitivity in both databases.

Therefore, these results point towards a higher relevance of the plosive segments of speech of parkinsonian patients in the automatic detection using speech.

The best AUC obtained in the first approach is 0.94, corresponding to an accuracy of $83 \pm 9\%$, employing the Neurovoz database and plosive distillation. In the second approach, the maximum AUC is 0.93 with an accuracy of $85 \pm 9\%$ in the same database.

Regarding the third approach, in which the allophonic distillation is applied only to the database used for the UBM, best AUC and accuracy are obtained employing plosives, vowels and fricatives with fixed texts and monologues when the adaptation-testing database is GITA. In this database, nasals and vowels distillation in Albayzin provide the best results when employing a DDK task, followed tightly by plosives.

Figure 8.9: ROCCH-DET curves for the three parkinsonian speech databases used for testing in the cross-database trials, employing the third approach (allophonic distillation only in UBM, Albayzin.) In all cases, DDK tasks from the parkinsonian databases are used.

Something different occurs in the Neurovoz database where fricative allophonic distillation in the UBM always produce the better AUC, accuracy and sensitivity. Plosives and vowels produce good results too, depending on the acoustic material employed.

In this third approach, best results are achieved employing DDK as speech task and fricative distillation in Albayzin while training and testing with CzechPD. These correspond to a $94 \pm 1\%$ of accuracy and an AUC of 0.96.

It is remarkable the fact that, although the DDK tasks used in this thesis do not include nasals or liquids, the nasal and liquid distillation of the database employed to obtain the UBM provide results comparable to the other distillations. The reasons for that are unclear but can be attributed to the inclusion of portions of vowels in the transitions of the nasal and liquid segments obtained with the speech forced alignment and that are used to train the UBM. Other reason is that the derivatives used as features consider portions of speech that go beyond the nasal

Figure 8.10: Best accuracies for the three different approaches (marked as 1, 2 or 3) and speech tasks, where ft stands for fixed text, ddk, Diadochokinetic task and monol., monologues.

or liquid selected segment, including information of articulation of the transition of vowels or other consonants. This information, contained in the [UBM](#) is adapted too using the [DDK](#) tasks.

Figures [9.11](#) and [9.12](#), in which the accuracies and [AUC](#) graphs can be observed, suggest that plosive distillation has a similar behavior in GITA and Neurovoz and provides high accuracy and [AUC](#) in most of the approaches. The other distillations have unequal results. For instance, while fricative distillation outperforms the other distillations in Neurovoz using the third approach, it does not provide these good results in GITA.

This trend cannot be confirmed with the [ROCCH-DET](#) curves of Figure [8.8](#), that depicts these curves obtained from the scores of the third approach, since all curves are relatively close and interlaced in all the three databases. However, the curves of figure [8.9](#), obtained after the cross-database trials, suggest a better performance of vowel segments, specially in the trials in which Neurovoz or CzechPD are used as testing database.

The fusion of scores using logistic regression produces some improvements in the trials with Neurovoz where the maximum [AUC](#) goes from 0.93 to 0.95 for fixed text and in CzechPD where the [AUC](#) goes from 0.96 to 0.98 using [DDK](#). In the rest of the cases, the fusion does not change the maximum values of accuracy and [AUC](#) or these are

Figure 8.11: Best AUC for the three different approaches (marked as 1, 2 or 3) and speech tasks, where ft stands for fixed text, ddk, Diadochokinetic task and monol., monologues.

reduced. However, although introducing the fusion does not produce the same effects in all the databases or speech tasks, the best results in terms of [AUC](#) and accuracy of the experimental set are obtained after this fusion.

Regarding the cross-database trials, whose results are shown in Table 8.9, although the accuracy and [AUC](#) are, generally, lower than in the rest of the trials, the sensibility is maintained or even increased. 0.86 values are obtained using GITA as testing database with plosive and nasal distillation, 0.93 and 0.91 with plosive and vowel distillations, respectively, in Neurovoz and 0.95 with nasal and plosive distillations in CzechPD. Again, plosive distillation is present in the three cited cases.

Also, although Czech and Neurovoz are class-unbalanced. The observed sensitivity and specificity rarely differ more than 0.10 absolute points in the best result cases.

Thus, results suggest that plosive segments tend to provide better results with the proposed schemes for the detection of [PD](#) most of the times, followed by vowels and fricative segments.

8.5.2 Analysis of approaches

In general, although the first and second approach help revealing which type of allophonic group segments can be more relevant for the detec-

tion of PD in the described approaches, the third approach outperforms the other two in terms of best accuracies, AUC and sensitivities. Although, ROCCH-DET curves of Figure 8.7 do not allow to extract any conclusion about which approach can be more advisable for a clinical environment, tables 8.3, 8.4, 8.5 and 8.6 suggest that the third approach tend to provide increased accuracy, AUC and sensitivity for all of the distillations than the other two approaches at the EER operating point. Also, ROCCH-DET curves in Figure 8.7 suggest that for other operating points providing a higher sensibility, which is desirable in clinical environments, this third approach is the most reliable. For these operating points, in GITA (a), this approach provides a similar accuracy than the first one, while the second under-performs the other two. With Neurovoz (b), something similar happens but, in this case, approach two is the one that under-performs respect the other two. Thus, these curves suggest that the third approach is the only one which provides a good performance for operating points providing high sensibility. Additionally, this approach does not require to perform forced alignment in the parkinsonian (adaptation-testing) databases and, therefore, it's computational cost is lower.

Moreover, the fact that the database employed to obtain the UBM is distilled helps to obtain models more precise for the distillation segments. For instance, after distilling Albayzin to obtain only fricatives, that are used to create the UBM, the final model, adapted from the UBM employing the training data from a parkinsonian database, will be more focused in the features of fricatives but will consider all the segments from the speech of patients and controls. This means that in the third approach, unlike in the other two, no segment is discarded and all can contribute with complementary information to differentiate between the two classes.

8.5.3 Speech tasks

Although the best results of this experimental set are obtained with a DDK task, this occurs in the CzechPD database, where the other speech tasks cannot be tested for this experimental set. Focusing in the other two parkinsonian databases, fixed texts always provide the best accuracy and AUC results. The reason for that lies in the fact that fixed texts contain more speech variability than DDK tasks and, at the same time, create more enclosed and precise models than monologues, as the type and number of phonemes in the training and testing utterances are always the same, allowing a better comparison between training and testing subjects. In other words, the obtained models are speech dependent, which imply a supervision but that can be more accurate than the unsupervised version of the approach, using monologues instead of fixed texts. On the other side, DDK tasks tend to provide better sensibility than the rest of the tasks in the two databases. Lastly, monologues always yield better correlation between scores and UPDRS,

reaching values up to 0.43 in the case of GITA and 0.52 with Neurovoz, suggesting that these speech tasks can be more appropriate to predict this rating scale.

8.5.4 *Other considerations*

It is possible to observe that among the two types of sonorant segments, vowels and nasals, vowels provide much better results. Considering that the amount of nasal segments is large enough to train the models (the number of nasal allophones in the fixed texts is similar to the other types of consonants), the better results provided by the vowel segments evidence that the proposed methodology is capturing articulatory characteristics related to PD during vowels production that are, obviously, not present in the sonorant consonants. The differences between patients and controls can, probably, be related to a different VSA caused by an incomplete articulation of the vowels which is characterized by the PLP features. In the same way, the nasal segments, after their parameterization, provide information about the nasal structures which may be not affected by PD in the same manner. In both types of segments, there is a common information about the source but, in any case, it is not well characterized by the selected features. The use of vowel segments leads to approaches that are comparing the formants (and, indirectly, the VSA) of a test speaker with the ones included in the models of the two classes, present in the Gaussian mixtures. At the same time, the derivatives of the Rasta-PLP characterizing the vowels are used as information about the velocity and acceleration of articulation of the speaker. Therefore, this approach employing the allophonic distillation of the vowels is taking advantage of certain particularities of the parkinsonian speech that in previous works were proven as successful.

8.6 SUMMARY

In this experimental set, a new concept was introduced: allophonic distillation, consisting in a frame selection depending on the manner of articulation, which can be fricative, liquid, nasal, plosive or vowel. To apply this distillation, speech forced alignment techniques were employed. Three different approaches to detect PD from speech, based on GMM-UBM schemes were presented. The differences between these approaches rely on the database to which the allophonic distillation is applied: the adaptation-testing database (parkinsonian), the UBM database or both. Observing the results, a scheme based on allophonic distillation only in the UBM database and employing fixed text as speaking task provides the best results. Regarding the type of distillation, results suggest that plosive, vowel and fricative (in this order)

allophonic distillation give the best results and a fusion of scores of different distillation maintains or increases these results.

Cross-validation trials (k-folds) provide accuracies between 85% and 94%, with AUC between 0.91 and 0.98 depending on the database, while cross-database trials provide accuracies between 75% and 82% with AUC between 0.84 and 0.95.

EXPERIMENTAL SET 4: ACOUSTIC LANDMARKS

9.1 INTRODUCTION

In the two previous chapters, the discriminative properties of certain phonetic segments of the speech is studied for the automatic detection of PD. These segments belonged to a certain phonetic category (/a/, /b/, /d/, etc.) or to a phonetic group attending to the manner of articulation (plosive, fricative, nasal, etc). The categorization of the segments was performed by means of speech forced alignment techniques which derives, most of the times, in the creation of supervised schemes in which the speaker must pronounce a fixed text. In this chapter another approach is proposed, using only transitions such as bursts or beginning or ending of glottal excitation which are selected after an acoustic analysis of the signal, producing unsupervised schemes.

Hence, in the experimental set included in this chapter, some specific points of speech, called Acoustic Landmarks [197], are detected in different speech tasks and the Rasta-PLP features around them are employed to detect the presence of PD through GMM and GMM-UBM classification techniques using several databases. The motivation of the use of acoustic landmarks is that these points coincide with very specific articulatory movements and the analysis of the signal around them could be of special interest for the detection of the influence of PD on speech. Consequently, this experimental set shares some methodological premises with the previous one, but in this case, the speech transcriptions are not necessary.

The use of acoustic landmarks for PD detection has been employed in [232] to extract prosodic features using these points as a basis. Other works like [122], [233] have used specific segments of speech like transitions to characterize and detect parkinsonian speakers.

In this experimental set, a new methodology is proposed in which the acoustic landmarks are employed to apply a distillation in the parkinsonian or in the auxiliary databases in a similar manner as in the previous experimental set. However, in this case, instead of using phonetic segments, as fricatives or plosives, only the speech signal around the acoustic landmarks are selected, containing only the transitions between allophones. Consequently, three different approaches for the automatic detection of PD are assessed, depending on which database the distillation is applied too.

9.2 THEORETICAL BACKGROUND

9.2.1 Acoustic landmarks

Acoustic Landmarks were first defined by Stevens in [197] as “a discrete representation of the speech stream in terms of a sequence of segments, each of which is described by a set (or bundle) of binary distinctive features”. In this experimental set, three types of landmarks are considered:

1. b-lmk, which are related to bursts during articulation (most of the times, the beginning of plosive consonants) or transitions between silences and fricatives or vice versa.
2. g-lmk, coinciding with the beginning or ending of vocal folds vibration.
3. s-lmk, transition point between vowels and sonorant consonants.

Figure 9.1 shows the spectrogram of a normal voice during the pronunciation of the sentence "how do we define it?"

Figure 9.1: Spectrogram and associated landmarks in the pronunciation of the sentence "how do we define it?". Red lines represent b-lmk, light and dark blue, g-lmk (pointing the beginning and end of vocal fold vibration, respectively) and green lines, s-lmk. Additionally, black lines mark the vowel landmark, in the middle of the two g-lmk and grey lines, the presence of a glide. The use of the signs +/− indicates that the transition is referred to the beginning (+) or the end (−) of an allophone. Obtained from [234].

Landmarks can be determined following the procedure described in [234], in which the energy changes in six frequency bands are analyzed so as to decide which type of landmark is detected. These bands, detailed in Table 9.1 are referred to the glottal vibration, sonorant segments or frication noise.

In this detection procedure, the abruptness on each band is calculated as the difference in energy between two speech frames separated a certain fixed time period. In a first coarse pass, the signal is divided into 16 ms frames and the abruptness is calculated as the difference in energy between two frames whose beginnings are separated 40 ms.

Table 9.1: Different frequency bands in landmark detection

Band	Frequency (Hz)	Description
1	0-400	Glottal vibration
2	800-1500	Zeros of sonorant
3	1200-2000	
4	2000-3500	Frication noise
5	3500-5000	
6	5000-8000	

The zones in which the difference is over a 7 dB threshold are marked as possible candidates to include a landmark. In a second pass, the same process is repeated but using 8 ms frames, time difference of 20 ms and threshold of 5 dB. All the abruptness values over this threshold are considered abruptness peaks. The points where the peaks from this fine pass coincide with a possible candidate from the coarse pass are selected as landmark candidates. The purpose of the coarse pass abruptness calculation is to avoid detecting abruptness caused by noise while the object of the fine pass is to locate more precisely the point of a candidate. If several peaks are found in the fine pass near a possible candidate in the coarse pass, only the one with highest abruptness is considered. Band 1 from Table 9.1 is used to detect the g-lmk preliminary candidates while bands from 2 to 6 are considered to detect b-lmk and s-lmk. When more than three peaks are found in these bands within a 50 ms interval, a b-lmk and s-lmk preliminary candidates are set. Figure 9.2 depicts this process. In order to decide the definitive candidate type between b-lmk and s-lmk, a statistical analysis of the acoustic conditions around the preliminary detected candidate is performed using the landmark acoustic cues.

Each landmark type (b-lmk, g-lmk and s-lmk) has several associated specific acoustic cues which consist on some measurements over the speech signal around the acoustic landmark. For b-lmk, the acoustic cues are abruptness (i.e., difference of energy level between two points separated by a certain time frame) and silence (i.e., energy level on both sides of the landmark). For g-lmk, acoustic cues are abruptness and vocalic level (again, energy level on both sides of the landmark). For s-lmk, the acoustic cues used in this work are abruptness and energy statistics (as mean, maximum, minimum and tilt). Then, the acoustic cues of each preliminary candidate b-lmk and s-lmk are introduced into a GMM classification system previously trained with a phonetically labeled database as explained in [234]. The system, contains the GMMs referred to each landmark type, allowing to calculate the probability of a candidate of being a true landmark (Probability of

Figure 9.2: Landmark preliminary candidate detection scheme based on the abruptness calculation.

Candidate (PoC)), which can be used, mainly, to differentiate between b-lmk and s-lmk.

9.2.2 Acoustic landmark distillation

Once the landmarks are detected, for a certain utterance, the signal can be distilled to obtain only the features (PLP, for instance) around certain frames of the landmarks and discarding the rest of the signals. Each frame produces, then, a feature vector. This process has been coined in this thesis as acoustic landmark distillation and in this experimental set is used to distill the parkinsonian and/or the UBM databases. The outcome of the acoustic landmark distillation is the features of all of the detected speech transitions.

9.3 EXPERIMENTAL SETUP

In this experimental set, the acoustic landmarks were detected for all of the Albayzin utterances, for fixed texts, DDK tasks and monologues from GITA and Neurovoz and for the DDK task in CzechPD. The landmark preliminary candidates, acoustic cues and PoC were calculated using a Matlab toolbox created by the Research Laboratory of Electronics at Massachusetts Institute of Technology¹ and detailed in [234]. To obtain the PoC, the GMM classification system was trained using the recordings and phonetic labels of Albayzin database. These labels were obtained through forced alignment as described in the two previous

¹ This toolbox is not available online or upon request and has been provided for this experimental set by the Research Laboratory of Electronics within a framework of collaboration between that group and the Bioengineering and Optoelectronic Laboratory of Universidad Politcnica de Madrid

experimental sets. All the possible candidate landmarks with a PoC under 0.1 are automatically discarded.

Firstly, a preliminary analysis of PoC using fixed texts from GITA database was performed, to evaluate the statistical behavior of the differences in the detection of landmarks in the parkinsonian and control classes for the same utterances. Considering that all the speakers pronounced the same texts within a certain database, the differences in the probability density functions will reveal if a group tends to produce a type of landmarks with more or less probability.

Then, three different approaches were considered. For the three approaches, $GMM-UBM$ classification techniques were employed in which a parkinsonian database (GITA, Neurovoz or CzechPD) adapts the UBM model trained with Albayzin. The difference between each approach refers to which speech database the acoustic landmark distillation is applied to. This experimental set is similar to the previous one in Chapter 8 (in which allophonic distillation techniques were considered) but in this case, instead of using whole phonetic segments, only the speech signal around the detected candidate landmarks is employed. In the proposed methodology, the acoustic landmark distillation resulted into three speech frames of 15 ms per acoustic landmark candidate detected, with a 50% overlap, as shown in Figure 9.3. Each frame produced a feature vector of Rasta- $PLP + \Delta + \Delta\Delta$.

Figure 9.3: Segments selected after acoustic landmark distillation process. When a candidate landmark is detected, only the features from the frame containing it and the two adjacent are employed in the PD detector.

In the first approach, the acoustic landmark distillation is applied only to the parkinsonian database, as depicted in Figure 9.4. In this manner, the information used to adapt the UBM as explained in section 6.2.3 and to test the resulting Gaussian model will contain only features from segments around specific transitions (b-, g- or s- lmk). In the second approach, the distillation is applied in the parkinsonian and in the UBM databases, as shown in 9.5, meaning that only specific transitions associated to the landmarks are used to train and test the classifiers. Finally, in the third approach this distillation procedure was applied only to the UBM , as shown in Figure 9.6, and the features from

whole utterances were employed to adapt the UBM and to test the classifier. In this manner, the obtained models were influenced by specific segments such as transitions between vowels and consonants or bursts but yet contained information about other articulation aspects of the speakers.

All the training-testing iterations on each approach follow a k -folds validation scheme with $k = 11$. In the same sense, the number of Gaussians of the GMM models are varied in the range $\{4, 8, 16, 32, 64, 128, 256\}$ while the number of Rasta-PLP coefficients was within the range $\{10, \dots, 20\}$ in steps of 2.

Figure 9.4: First approach for PD detection based on the detection of landmarks.

Figure 9.5: Second approach for PD detection based on the detection of landmarks.

Figure 9.6: Third approach for PD detection based on the detection of landmarks.

Finally, and considering only the approach that led to best results, a fusion of scores employing logistic regression was performed for all of the $n - tuples$ combinations of the scores resulting from the b-lmk, g-lmk and s-lmk distillations, going from 2 – tuples to 3 – tuples. Finally, a cross-database validation was carried out contemplating only the approach leading to the best results. In this type of validation process the **DDK** tasks from two out of the three available parkinsonian databases, were employed to adapt the **UBM** model while the remaining was used to test the classifier. This process was repeated three times, leaving at each round, a different parkinsonian database as testing set.

9.4 RESULTS

This section includes the tables with the results in terms of accuracy, **CI**, **AUC**, specificity, sensitivity and correlation of scores with **UPDRS** and **H-Y** along with their respective p-values obtained with the three different approaches. Table 9.2 contains the results of the first approach in which the acoustic landmarks are detected (and utilized) only in the adaptation-testing (parkinsonian speech) databases. Table 9.3 contains the results of the second approach, in which the landmarks are detected for adaptation-testing and **UBM** databases. Figures 9.7 and 9.8 include the probability density functions of the **PoC** associated to the three types of landmarks in GITA and Neurovoz respectively.

Table 9.4 contains the results of the third approach in which the landmarks are detected only in the **UBM** database. In the three cases, the results of fixed texts, monologues and **DDK** are included for GITA and Neurovoz and only **DDK** tasks are employed in the CzechPD database in order to use Albayzin as **UBM**. As this approach leads to the higher accuracy and **AUC** values than the previous ones, the associated **ROCCH**-

DET curves obtained with the three databases are included in Figure 9.9. Fixed text was selected for these curves in GITA and Neurovoz since this task provides the best results.

Table 9.5 includes the results of the fusion of scores in the third approach while Table 9.6 provides the results of the crossdatabase tests in which two parkinsonian databases out of the three available, train the Gaussian models employing Albayzin as UBM while the remaining one performs the testing of the models. In these two last cases, only the third approach is considered as is the one leading to better results.

Table 9.2: First approach results for the three types of landmarks and databases using Albayzin for the UBM.

Database	Speech task	lmk	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed Text	b	80 $\pm$ 8	0.83	0.76	0.84	12	16	0.16	0.2911	0.19	0.1975
		g	79 $\pm$ 8	0.84	0.64	0.94	14	128	0.16	0.284	0.16	0.2714
		s	78 $\pm$ 8	0.87	0.7	0.86	12	64	0.05	0.724	0.04	0.7749
	DDK	b	75 $\pm$ 8	0.84	0.74	0.76	20	64	-0.01	0.9671	-0.01	0.9251
		g	73 $\pm$ 9	0.84	0.76	0.7	10	8	0.06	0.6909	-0.06	0.6957
		s	70 $\pm$ 9	0.79	0.78	0.62	10	32	0.02	0.8679	0.14	0.3367
	Monol.	b	76 $\pm$ 8	0.8	0.76	0.76	12	4	0.24	0.101	0.15	0.2881
		g	77 $\pm$ 8	0.83	0.69	0.84	18	64	0.39	0.0052	0.3	0.0358
		s	75 $\pm$ 9	0.83	0.71	0.78	10	32	0.36	0.0109	0.22	0.1235
Neurovoz	Fixed Text	b	89 $\pm$ 7	0.93	0.91	0.84	14	128	0.31	0.0706	0.05	0.7718
		g	80 $\pm$ 9	0.9	0.79	0.81	14	4	0.27	0.1184	0.18	0.2814
		s	75 $\pm$ 10	0.86	0.7	0.81	10	16	0.44	0.0089	0.43	0.0063
	DDK	b	81 $\pm$ 9	0.89	0.8	0.81	14	256	0.08	0.6611	-0.01	0.9459
		g	79 $\pm$ 9	0.81	0.83	0.73	12	32	0.15	0.3943	0.03	0.846
		s	79 $\pm$ 9	0.88	0.76	0.85	12	16	0.2	0.2596	0.26	0.1145
	Monol.	b	79 $\pm$ 12	0.9	0.47	0.97	14	16	0.33	0.2651	0.3	0.3152
		g	70 $\pm$ 13	0.71	0.47	0.83	18	16	0.54	0.0572	0.46	0.1101
		s	74 $\pm$ 12	0.75	0.59	0.83	10	8	0.41	0.1594	0.27	0.3814
CzechPD	DDK	b	88 $\pm$ 1	0.95	0.8	1	16	16	0.17	0.4736	0.15	0.5367
		g	88 $\pm$ 1	0.95	0.9	0.86	18	4	0.22	0.3472	0.27	0.2495
		s	79 $\pm$ 14	0.86	0.9	0.64	14	256	-0.13	0.5985	-0.06	0.8175

Figure 9.7: Probability density functions of landmarks' PoC for fixed texts in GITA.

Figure 9.8: Probability density functions of landmarks' PoC for fixed texts in Neurovoz.

Best results per database, speech tasks and approach are shaded in green and best global results per database are in bold.

Finally, Figure 9.10 shows the ROCCH-DET curves of the cited cross-database trials and Figures 9.11 and 9.12 include a graphical representation of highest accuracies and AUC reached in the different trials.

9.5 DISCUSSION

In this experimental set, in a similar manner as in the the experimental set 3, three approaches using GMM-UBM classification schemes are employed, utilizing acoustic landmark distillation in the parkinsonian, or the UBM databases or in both. In this case, these schemes allow to analyze the importance of the transitions during articulation, which are the moments in which the coordination of movements is, usually, more complex.

In general, as in previous experimental sets, the CI is over 7% in GITA and Neurovoz trials and, therefore, the analysis included in this section is based on the observed trends. In most of the cases the comparison of results cannot be conclusive as a consequence of the overlapping margins between results.

Table 9.3: Second approach results for the three types of landmarks and databases using Albayzin for the UBM.

Database	Speech task	lmk	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed Text	b	77 $\pm$ 8	0.85	0.66	0.88	10	64	0.08	0.5726	0.09	0.556
		g	81 $\pm$ 8	0.86	0.8	0.82	10	64	0.05	0.7587	0.06	0.7041
		s	77 $\pm$ 8	0.85	0.7	0.84	12	64	0.11	0.476	0.08	0.5741
	DDK	b	76 $\pm$ 8	0.82	0.72	0.8	12	128	-0.02	0.8954	-0.04	0.8036
		g	79 $\pm$ 8	0.85	0.78	0.8	10	4	-0.01	0.9289	-0.07	0.6386
		s	75 $\pm$ 8	0.82	0.8	0.7	10	16	0.23	0.1237	0.27	0.0691
	Monol.	b	77 $\pm$ 8	0.84	0.63	0.9	20	64	0.3	0.0366	0.3	0.0339
		g	78 $\pm$ 8	0.82	0.69	0.86	18	128	0.36	0.0123	0.26	0.0731
		s	75 $\pm$ 9	0.83	0.69	0.8	10	128	0.33	0.0203	0.28	0.0544
Neurovoz	Fixed Text	b	86 $\pm$ 8	0.93	0.81	0.94	12	128	0.25	0.1471	0.03	0.8612
		g	81 $\pm$ 9	0.9	0.81	0.81	12	64	0.16	0.3736	-0.11	0.5186
		s	77 $\pm$ 9	0.88	0.77	0.78	12	32	0.5	0.0024	0.33	0.0404
	DDK	b	82 $\pm$ 9	0.91	0.8	0.85	14	256	0.02	0.9272	-0.1	0.5437
		g	82 $\pm$ 9	0.86	0.87	0.73	16	256	0.25	0.1534	-0.05	0.7442
		s	82 $\pm$ 9	0.9	0.83	0.81	10	4	0.29	0.0972	0.31	0.0607
	Monol.	b	77 $\pm$ 12	0.85	0.41	0.97	18	16	0.49	0.0924	0.37	0.21
		g	70 $\pm$ 13	0.73	0.53	0.8	14	8	0.43	0.1429	0.4	0.1728
		s	72 $\pm$ 13	0.75	0.65	0.77	16	16	0.17	0.5677	0.08	0.7946
CzechPD	DDK	b	88 $\pm$ 1	0.97	0.8	1	20	32	0.01	0.9753	0.03	0.9135
		g	91 $\pm$ 1	0.94	0.85	1	16	32	0.08	0.7385	0.06	0.809
		s	82 $\pm$ 13	0.82	0.75	0.93	20	64	-0.03	0.8926	-0.07	0.7661

Table 9.4: Third approach results for the three types of landmarks and databases using Albayzin for the UBM.

Database	Speech task	lmk	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed Text	b	85 ± 7	0.89	0.84	0.86	12	128	0.13	0.3832	0.11	0.4456
		g	82 ± 8	0.89	0.8	0.84	12	64	0.1	0.485	0.06	0.673
		s	82 ± 8	0.89	0.78	0.86	12	128	0.13	0.3903	0.07	0.6551
	DDK	b	83 ± 7	0.88	0.86	0.8	10	8	-0.01	0.9258	-0.06	0.7134
		g	82 ± 8	0.88	0.78	0.86	16	128	0.01	0.9249	-0.03	0.8602
		s	83 ± 7	0.88	0.82	0.84	10	4	-0.04	0.8132	-0.06	0.7086
	Monol.	b	77 ± 8	0.87	0.67	0.86	20	256	0.29	0.0407	0.21	0.1436
		g	80 ± 8	0.86	0.76	0.84	12	64	0.38	0.0067	0.23	0.1129
		s	78 ± 8	0.87	0.73	0.82	12	128	0.4	0.0046	0.28	0.0552
Neurovoz	Fixed Text	b	87 ± 7	0.92	0.87	0.88	16	128	0.36	0.0334	0.05	0.7709
		g	85 ± 8	0.92	0.87	0.81	12	64	0.34	0.0483	0.05	0.7625
		s	89 ± 7	0.93	0.85	0.94	12	256	0.4	0.0163	0.17	0.303
	DDK	b	83 ± 9	0.91	0.85	0.81	10	64	0.14	0.4322	0.12	0.4576
		g	81 ± 9	0.89	0.8	0.81	10	256	0.1	0.5641	0.07	0.6609
		s	79 ± 9	0.88	0.83	0.73	10	256	0.15	0.3903	0.12	0.4823
	Monol.	b	70 ± 13	0.72	0.35	0.9	14	64	0.09	0.7816	-0.21	0.4872
		g	70 ± 13	0.74	0.53	0.8	18	32	0.45	0.1213	0.34	0.2577
		s	74 ± 12	0.79	0.47	0.9	16	64	0.44	0.1359	0.28	0.3531
CzechPD	DDK	b	94 ± 1	0.97	0.9	1	20	32	-0.01	0.9665	-0.04	0.8607
		g	94 ± 1	0.97	0.9	1	16	32	0.03	0.9074	0.06	0.8028
		s	94 ± 1	0.97	0.9	1	18	64	0.09	0.7154	0.06	0.7898

Table 9.5: Fusion of scores in the third approach for the three parkinsonian databases using Albayzin as UBM. For the sake of simplicity, the table includes only the combination of scores (2- and 3 - *tuples*) leading to the best results.

Database	Speech task	Combination	Accuracy $\pm$ CI	AUC	Sens.	Spec.	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	Fixed text	b-lmk-s-lmk	84 ± 7	0.89	0.82	0.86	0.11	0.4475	0.11	0.4782
	DDK	b-lmk-s-lmk	82 ± 8	0.87	0.8	0.84	0.04	0.8071	0.02	0.8742
	Monol.	b-lmk-g-lmk	80 ± 8	0.87	0.76	0.84	0.37	0.009	0.25	0.0872
Neurovoz	Fixed text	b-lmk-s-lmk	86 ± 8	0.95	0.81	0.94	0.35	0.0367	0.33	0.0375
	DDK	b-lmk-g-lmk-s-lmk	83 ± 9	0.88	0.87	0.77	0.26	0.1351	0.24	0.1521
	Monol.	b-lmk-g-lmk	72 ± 13	0.75	0.65	0.77	0.35	0.2344	0.36	0.2237
CzechPD	DDK	b-lmk-s-lmk	94 ± 1	0.99	0.9	1	0.27	0.2473	0.12	0.6062

Table 9.6: Cross-database results using DDK task and the third approach (in which the landmarks are only used in the UBM database. For every trial, two parkinsonian databases are used for training and the remaining, for testing.

Test database	Speech task	lmk	Accuracy $\pm$ CI	AUC	Sens.	Spec.	N	G	UPDRS ρ	p-val.	H-Y ρ	p-val.
GITA	DDK	b	73 ± 9	0.8	0.72	0.74	18	32	0.02	0.8953	-0.07	0.645
		g	71 ± 9	0.77	0.94	0.48	18	128	0.04	0.8134	-0.07	0.6473
		s	73 ± 9	0.81	0.8	0.66	18	256	0.12	0.4049	0.04	0.8023
Neurovoz	DDK	b	74 ± 10	0.81	0.87	0.5	10	4	0.2	0.2598	0.13	0.4463
		g	74 ± 10	0.8	0.83	0.58	16	8	0.3	0.0871	0.32	0.0501
		s	76 ± 10	0.83	0.78	0.73	20	64	0.46	0.0068	0.31	0.0558
CzechPD	DDK	b	76 ± 14	0.91	0.95	0.5	10	16	0.18	0.4469	0.16	0.4876
		g	85 ± 12	0.94	0.95	0.71	14	8	0.09	0.6934	0.08	0.7362
		s	82 ± 13	0.95	1	0.57	12	4	0.25	0.2822	0.17	0.4629

Figure 9.9: ROCCH-DET curves for the three parkinsonian speech databases in the third approach (acoustic landmark detection used only on UBM database, Albayzin). Fixed text are used for GITA and Neurovoz databases, and DDK for CzechPD.

9.5.1 Types of landmarks

Initially, the analysis of Figures 9.7 and 9.8 reveals that in both databases, the density of probability of **PoC** in b-lmk is lower for higher values of probability in patients than in controls. Therefore, the **GMM** is modelling the probability of b-lmk is scoring the b-lmk of patients with lower values than those of controls, suggesting that the bursts of patients are not as expected or, at least, not as the typical bursts. Considering that all speakers are pronouncing the same fixed texts within a certain database, all are expected to have the same amount of bursts. These differences in the probability density functions of the b-lmk can be caused by the motor perturbation called spinantization associated to articulation that many **PD** patients suffer in which some burst or plosive consonants tend to turn into fricative or are completely transformed by the adjacent sonorant. This sign may be a consequence of the reduction of the articulation ranges. These differences between the

Figure 9.10: ROCCH-DET curves for the three parkinsonian speech databases used as testing set in a cross-database scheme employing the third approach (acoustic landmark detection used only on UBM database, Albayzin). Only DDK tasks are utilized.

two classes are unclear in the cases of the other two types of landmarks.

Regarding the first approach, the highest accuracies, [AUC](#) and sensibility are obtained with b-lmk, as it can be inferred from [Table 9.2](#). The differences between the landmark type are more pronounced in the case of Neurovoz. However, in the second approach in which the landmark distillation is applied in the parkinsonian databases and in Albayzin ([UBM](#)) b-lmk produce the best results in Neurovoz and g-lmk, in GITA and CzechPD. In all the cases, the differences between the results obtained using the different landmark types are less pronounced than in the first approach. In the third approach, in which the landmark distillation is applied only to Albayzin to obtain the [UBM](#) (results shown in [Table 9.4](#)), b-lmk provide the best results using fixed text in GITA while in Neurovoz, s-lmk produce the best results employing the same speech task. For the rest of the tasks, b-lmk and s-lmk have a similar performance. Regarding CzechPD, all three types of landmarks

Figure 9.11: Best accuracies for the three different approaches (marked as 1, 2 or 3) and speech tasks, where ft stands for fixed text, ddk, Diadochokinetic task and monol., monologues.

Figure 9.12: Best AUC for the three different approaches (marked as 1, 2 or 3) and speech tasks, where ft stands for fixed text, ddk, Diadochokinetic task and monol., monologues.

produce the best results. In any case, results obtained employing the three types of landmarks are, in general, close, with lower accuracies when employing g-lmk, effect that can be observed in Figure 9.9. In these ROCCH-DET curves it is visible that the three curves are interlaced although g-lmk distillation tends to provide lower global results, specially for the EER operating point.

The fusion of scores of the third approach does not provide substantial improvements, attending to the results shown in Table 9.5 although higher AUC in Neurovoz and CzechPD are observed. This can be explained by the fact that after the fusion, the ROC curve is smoothed and the AUC improves. However, it is noticeable that in all the cases b-lmk are involved in the fusion providing the best results together with s-lmk, in most of the cases.

Although, in general, results suggest that there exist a trend to obtain a better performance of the systems when b-lmk are used for the distillation of the UBM, cross-database point out to a better performance of s-lmk, followed closely by b-lmk, as shown in Table 9.6.

These observations are supported by the graphs included in Figures 9.11 and 9.12, in which b-lmk tend to provide higher accuracies and AUC in most of the databases, approaches and speaking tasks, followed closely by s-lmk.

Nevertheless, it is important to mention that for operating points providing higher sensibilities, s-lmk tend to have a better accuracy, as suggested by Figures 9.9 and 9.10.

9.5.2 Analysis of approaches and speech tasks

In this experimental set, similarly to the experimental set 3, best results are obtained using landmark distillation in the UBM only, as deduced from Tables 9.2, 9.3 and 9.4. Best results employing GITA database entail the use of fixed text and the use of b-lmk (85% of accuracy and 0.89 of AUC). However, in the case of Neurovoz, best results are reached using s-lmk and fixed texts (89% and 0.93).

In general, best results are obtained with text-dependent systems as in the previous experimental sets.

Lastly, and regarding the cross-database trials performed employing the third approach, the maximum AUC obtained were 0.81, 0.83 and 0.85 for GITA, Neurovoz and CzechPD, respectively. It is important to remark that although the accuracies of the cross-database trials are lower than those obtained with the k-folds cross validation, the sensibilities of the systems are similar in both cases. As in previous experiments, an increase of the accuracy can be expected by using fixed texts in cross-database trials but, in this thesis, it is not possible to perform these trials since the three databases contain different types of fixed texts.

9.5.3 *Other considerations*

In this experimental set, the minimum PoC required to determine if a landmark is a true landmark has not been studied. Most of the detected landmark candidates have been used since it has been considered that if a point fulfills the conditions to be a landmark, that point could be of interest to detect PD since it could be indicating the existence of a point with articulatory information. In the future, new tests using only true landmarks, selected by combining the detection technique exposed here with HMM, Viterbi algorithms or, directly, speech forced alignment techniques, should be addressed.

9.6 SUMMARY

In this experimental set a new frame selection technique, acoustic landmarks distillation, is employed to use specific segments of speech like bursts or transitions between vowels and consonants, among others, to train and test new speech-based automatic detectors of PD employing GMM-UBM techniques.

Three different approaches are proposed, applying the distillation to different databases (adaptation-testing or auxiliary) and the results employing the three parkinsonian databases are analyzed along with a fusion of scores approach and cross-database validations. In general, the landmark distillation in the speech employed to train the UBM using b-lmk tends to provide better accuracies at EER operating points but, in order to obtain systems with higher sensibility, s-lmk distillation is more advisable, as suggested by results.

Cross-validation trials (k-folds) provide accuracies between 85% and 94%, with AUC between 0.89 and 0.99 depending on the database, while cross-database trials provide accuracies between 73% and 82% with AUC between 0.81 and 0.95.

Part III

PHONATORY APPROACHES

EXPERIMENTAL SET 5: PHONATORY FEATURES

After the proposal and analysis of several articulatory approaches, this chapter details a new experimental set exploring the properties of several characterization methods and classification schemes for the differential diagnosis of PD using sustained phonation (vowel /a:/) of speakers as input source to the system. In particular, four types of characterizations are used: noise, amplitude and frequency perturbations (jitter and shimmer), complexity measurements, PLP coefficients and Modulation Spectrum (MS) features. The first three types possess discriminating properties regarding the detection of PD, as literature has already proven. The fourth was tested in experimental set 1 with reasonably good results. The fifth one is a new type of parameterization proposed by the author of this thesis in order to characterize frequency and amplitude modulation perturbations in pathological voices.

The main purpose of these features is to measure the presence of noise, amplitude, frequency and dynamic (linear and non-linear) perturbations, that in the case of parkinsonian speakers are caused by the symmetric and non-symmetric rigidity of the muscles controlling the vocal folds, incoordination of movements in the phonatory system and tremor.

To this respect, and after a prospective observation of the spectrum, waveform and listening to the recordings of sustained vowels in the referred databases, a common pattern in many patients is the decreasing SPL produced during phonation and the presence of noise, specially at the end of the vowel. Figure 10.1 shows these effects in four patients. This noise can be justified not only by problems in the control of the larynx and phonatory muscles but in the muscles related to swallowing actions, which in some cases lead to the presence of saliva in the vocal folds. Figure 10.2 includes the waveform and spectrograms of four control speakers, allowing to compare cited parkinsonian effects with Figure 10.1.

(a) Idiopathic PD male speaker. Age: 88. UPDRS: 59. Span: 1,995 s

(b) PD female speaker. Age: 77. UPDRS: 35. Span: 2,995 s

(c) Idiopathic PD male speaker. Age: 82. UPDRS: 26. Span: 3,713 s

(d) PD female speaker. Age: 71. UPDRS: 12. Span: 3,136 s

Figure 10.1: Waveform and spectrogram of four patients during the sustained phonation of vowel /a:/. Obtained from Neurovoz database.

(a) Control male speaker. Age: 77. Span: 4,483 s

(b) Control female speaker. Age: 74. Span: 3,947 s

(c) Control male speaker. Age: 74. Span: 3,743 s

(d) Control female speaker. Age: 71. Span: 4,252 s

Figure 10.2: Waveform and spectrogram of four patients during the sustained phonation of vowel /a:/. Obtained from Neurovoz database.

10.1 THEORETICAL BACKGROUND

This section describes the theoretical concepts associated to the features used to characterize the voice signal in this experimental set, mainly those related to [MS](#) as these are new features proposed as part of this work.

10.1.1 *Jitter, shimmer, noise and complexity features*

In this experimental set, several features, traditionally used to measure frequency or amplitude perturbations and noise are used to detect dysfunctions on the phonation of patients. In this sense, complexity features, which have been explored in the literature more recently for the same purposes and yielding good results in the detection of pathological voices, are used too. All of these features are described in this section.

10.1.1.1 *Jitter and shimmer*

Jitter measurements are aimed to characterize the frequency perturbations of the glottal pulse [235], [236] by quantifying the differences of fundamental period between consecutive glottal cycles of the signal. To measure the jitter, the temporal points coinciding with a certain peak of each glottal cycle (normally, the maximum) must be determined. Once these points are calculated, the absolute mean jitter (*jitt*) for a certain signal is calculated as the mean difference between the period of each cycle and the period of the next cycle:

$$jitt = \sum_{i=1}^{B-1} \tau_{i+1} - \tau_i \quad (10.1)$$

where B is the total number of periods detected in a certain signal and τ_i is the period of the glottal cycle i .

Other common jitter measurement is the relative jitter *%jitt* which consists in the normalization of *jitt* by the mean period of the signal:

$$\%jitt = \frac{\sum_{i=1}^{B-1} \tau_{i+1} - \tau_i}{\frac{1}{B} \sum_{i=1}^B \tau_i} \quad (10.2)$$

In a similar manner, shimmer quantifies the differences between the amplitudes of the peaks of the different glottal cycles within a signal. To calculate shimmer features, these glottal peaks must be determined, as in the case of jitter. Now, to calculate the absolute mean shimmer

Figure 10.3: Jitter and shimmer phenomena reflected in a sustained vowel signal.

(*shim*), the mean differences between the amplitudes of adjacent cycles are computed:

$$shim = \sum_{i=1}^{B-1} Y_{i+1} - Y_i \quad (10.3)$$

being Y_i the maximum amplitude of the cycle i . In the same sense, the relative shimmer *%shim* is the *shim* normalized respect to the mean amplitude:

$$\%shim = \frac{\sum_{i=1}^{B-1} Y_{i+1} - Y_i}{\frac{1}{B} \sum_{i=1}^B Y_i} \quad (10.4)$$

Figure

10.1.1.2 Noise

Noise features applied to the quality of voice assessment, are mainly aimed to quantify the harmonic nature of the phonation.

In this experimental set, the noise features Normalized Noise Energy (*NNE*) [237], *HNR* [238] and Glottal to Noise Excitation ratio (*GNE*) [239] are employed.

The *NNE* is the ratio between the noise energy level and the total energy level of the signal during the phonation. To determine the noise level, the fundamental frequency of the signal is measured in order to identify all the harmonics in the spectrum of the signal. The noise energy level is estimated using the energy from the valleys between the harmonics.

The *HNR* is calculated in a similar manner but, in this case, this value represents the ratio between the harmonic and the noise components from voice, and, although it could be considered as the inverse of *NNE*,

the procedure to calculate it is different. The harmonic level is calculated, as in [NNE](#), using the whole signal. However, the noise energy is estimated by removing the harmonic energy from the signal in the cepstral domain with a filter (liftering). The cepstral frequency (frequency) of the harmonics (rahmonics) is determined by measuring the fundamental frequency beforehand.

Lastly, the [GNE](#) was proposed to characterize the turbulent noise of voice. It is defined as the maximum correlation between the Hilbert envelope of several frequency bands extracted from the glottal pulse, which is obtained using inverse filtering techniques. To obtain these bands through the use of a filterbank, the glottal source is estimated by means of an inverse filtering of the signal.

10.1.2 Modulation Spectrum

[MS](#) provides information about the energy at modulation frequencies that can be found in the carriers of a signal. It is a three-dimensional representation where abscissa represents modulation frequency, ordinate axis depicts acoustic frequency and applicate, acoustic energy. This kind of representation allows observing different voice features simultaneously such as the harmonic nature of the signal and the modulations present at fundamental frequency and its harmonics. For instance, the presence of tremor, understood as low frequency perturbations of the fundamental frequency, can be easily noticeable since it implies a modulation of fundamental frequency as an usual effect of laryngeal muscles improper activity. Other modulations associated to fundamental or harmonic frequencies could indicate the presence of a dysfunction of the phonatory system.

To obtain [MS](#), the signal is filtered using a short-Time Fourier Transform ([sTFT](#)) filter bank whose output is used to detect amplitude and envelope. This output is analyzed using Fast Fourier Transform ([FFT](#)) producing a matrix E where [MS](#) values at any point can be represented as $E(f_a, f_m)$. The columns at E (fixed f_m) are modulation-frequency bands, and rows (fixed f_a) are acoustic-frequency bands. Therefore, a can be interpreted as the number of acoustic band and m , the number of modulation band while f_a and f_m are the central frequencies of the respective bands. Due to the fact that values $E(f_a, f_m)$ have real and imaginary parts, the original matrix can be represented using the modulus $|E|$ and the phase $\arg(E)$ of the spectrum. Throughout this experimental set, the [MS](#) has been calculated using the Modulation Toolbox library ver 2.1 [240]. Some different configurations can be used to obtain E , where the most significant degrees of freedom are the use of coherent or non-coherent (Hilbert envelope) [241] modulation, the number of acoustic bands, and acoustic and modulation frequency ranges.

Figure 10.4 shows an example of [MS](#) extracted from two different voices on which it is possible to see that the energy of the patient's

voice is more dispersed in comparison to the energy of the control speaker, in which case it is more concentrated in the modulation band referred to 0 Hz. This is a clear example on how voice perturbations can affect MS. This type of representation can reflect the presence of tremor as an increase of the energy in the low modulation frequencies around fundamental frequency. In the same manner, the uneven vibration of the vocal folds is reflected in the MS as a dispersion of the energy concentrated in the acoustic axis, with 0 Hz of modulation frequency to higher modulation frequencies. Equally, the presence of turbulences and noise is reflected as the dispersion of the energy concentrated in the harmonics, specially in mid-high frequency, leading to a smoother MS in the harmonics zone.

(a) Idiopathic PD male speaker. Age: 88. UPDRS: 12. (b) Control male speaker. Age: 74.

(c) Idiopathic PD newly diagnosed female speaker. Age: 80. UPDRS: 5. (d) Control female speaker. Age: 79.

Figure 10.4: MS of a two patient and two control speakers during the sustained phonation of vowel /a:/. Obtained from Neurovoz database.

Despite of its advantages, one of the principal drawbacks of MS is that it provides a large amount of information that can not be easily processed automatically. In this sense, MS matrices have to be processed to obtain a more compact but precise enough representation of the speech segments. Thus, after obtaining the MS, some representative features are extracted to feed a further classification stage. With this in mind, a group of features based on MS are proposed in this thesis¹:

¹ These were included in the works [242], [243]

Low Modulation Ratio (**LMR**), dispersion features (Cumulative Intersection Level (**CIL**) and Ratio of Points Above Linear Average (**RALA**) and Modulation Spectrum Percentiles (**MSP**)), Modulation Spectrum Contrast (**MSW**) and Modulation Spectrum Homogeneity (**MSH**). All of these features use the **MS** modulus as input source, except the last two which also use the phase.

10.1.2.1 Low Modulation Ratio

LMR, expressed in dB, is the ratio between energy in the first modulation band $\varepsilon(f_{a(f_0)}, f_1)$ at fundamental frequency f_0 and the global energy in all modulation bands covering at least from 0 to 25 Hz at acoustic frequency f_0 , $\varepsilon(f_{a(f_0)}, f_{m(25\text{Hz})})$. Its calculation is carried out according to the expressions (10.5) and (10.6). These bands are represented in Figure 10.5.

$$LMR = 10 \cdot \log \left(\frac{\varepsilon(f_{a(f_0)}, f_1)}{\varepsilon(f_{a(f_0)}, f_{m(25\text{Hz})})} \right) \quad (10.5)$$

being

$$\varepsilon(f_a, f_k) = \sum_{m=1}^k |E(f_a, f_m)|^2 \quad (10.6)$$

where $a(f_0)$ is the number of the acoustic band including fundamental frequency and $m(25 \text{ Hz})$, the number of the modulation band including 25 Hz.

The **MS** points used to calculate **LMR** are represented in Figure 10.5.

The 0-25 Hz band has been selected to represent all possible cases of tremor and low frequency modulations around pitch frequency [244], [245].

10.1.2.2 Contrast and Homogeneity.

Representing **MS** (modulus) as two dimensional images allows to observe that pathological voices usually have more complex distributions. Images related to normal voices are frequently more homogeneous and present less contrast, as can be seen in Fig. 10.4. Accordingly, Homogeneity and Contrast are used as two **MS** features since they provide information about the existence of voice perturbations.

Homogeneity is computed using the Bhanu method described by equation (10.7), as stated in [246].

$$MSH = \sum_a \sum_m [E(f_a, f_m) - \overline{E(f_a, f_m)}_{3 \times 3}]^2, \quad (10.7)$$

being MSH the **MS** Homogeneity value; $E(f_a, f_m)$ the modulation spectra computation (modulus only) at point (f_a, f_m) ; and $\overline{E(f_a, f_m)}_{3 \times 3}$ the average value in a 3×3 window centered at the same point.

Figure 10.5: Points of the MS matrix used to obtain MSH (i), MSW (ii) and LMR (iii).

Contrast is computed using a variation of the Weber-Fechner contrast relationship method described by equation (10.8) as stated in [246].

$$MSW(f_a, f_m) = \sum_{a'} \sum_{m'} C_{f_a, f_m} \quad (10.8)$$

where

$$C_{f_a, f_m} = \frac{|E(f_a, f_m) - E(f_{a'}, f_{m'})|}{|E(f_a, f_m) + E(f_{a'}, f_{m'})|} \quad (10.9)$$

representing $(f_{a'}, f_{m'})$ the vertical and horizontal adjacent points to (f_a, f_m) . The global MSW is considered as the sum of all points in $MSW(f_m, f_a)$ divided by the total number of points to normalize.

The MS regions used to calculate MSH and MSW at each point of the matrix are represented in Figure 10.5.

10.1.2.3 Dispersion features

As MS differs from normal to pathological voices, changes in the histograms of MS modulus reflect the effects of a dysfunction in a patient's voice. A short view to the MS allows to observe that many parkinsonian voices usually have a larger number of points with levels above the average value of $|E(f_m, f_a)|$. The presence of energy at modulation frequencies beyond the central modulation band (0 Hz) can be interpreted as a dispersion, which, in turns, is caused by anomalies in the glottal source functioning. Figure 10.4 depicts two typical cases of

the differences on dispersion between normal and parkinsonian speakers. In this figure, it is possible to observe that in the case of patients, the energy is more spread in the modulation bands than in the case of controls.

With this in mind, three dispersion features are proposed to measure such dispersion effect: **CIL**, **RALA** and **MSP**. **CIL** is the intersection between the increasing and decreasing histogram cumulative curves. Histogram is processed from **MS** modulus in logarithmic units (dB). On the other hand, **RALA** represents the ratio of points in **MS** modulus which are over the average and the number of points which are above this average in $E(f_m, f_a)$. Calculation of **RALA** is detailed in the expression (10.13).

$$RALA = \frac{NA}{NB} \quad (10.10)$$

being

$$NA = \sum_{f_a} \sum_{f_m} \omega(f_a, f_m) \quad (10.11)$$

$$NB = \sum_{f_a} \sum_{f_m} 1 - \omega(f_a, f_m) \quad (10.12)$$

and

$$\omega(f_a, f_m) = \begin{cases} 1 & |E(f_a, f_m)| \geq \overline{|E|} \\ 0 & |E(f_a, f_m)| < \overline{|E|} \end{cases} \quad (10.13)$$

where $\overline{|E|}$ is the **MS** modulus average, NA , the number of points above $\overline{|E|}$, NB the number of points below $\overline{|E|}$ and NT the total number of points in $E(f_m, f_a)$.

Lastly, taking into consideration **MS** modulus, 25%, 75% and 95% percentiles are proposed as features. These statistical measurements, called **MSP25**, **MSP75** and **MSP95** respectively, point out the presence of energy on **MS** at different level ranges and serve as indicators of the level distribution at low, mid-low and high relative level ranges given that all the signals into the database used to calculate these values are normalized. **MSP95** indicates the value under which the lower level values are. **MSP75** indicates the value in which the mid-low levels are around and **MSP25** is referred to the mid-high levels. The higher the number of points of high levels in **MS**, the higher the **MSP25** value. The higher the number of points of low levels, the lower the **MSP95** value. By way of illustration, with these metrics it is possible to compare if a voice has more high-level or low-level points at a particular modulation frequency range than other.

10.2 METHODOLOGY

Four different approaches were analyzed in this experimental set, attending to the type of features used to characterize the sustained phonation of the speakers. In the first one, noise, jitter, shimmer and complexity measurements formed the front-end. In the second one, the [MS](#) features were employed while in the third one, Rasta-[PLP](#)+ Δ + $\Delta\Delta$ coefficients form the feature vectors. Lastly, in the fourth approach, all the cited features were used to train and test the detectors along with a feature dimensionality reduction technique (Principal Components Analysis ([PCA](#))) and a feature selection technique (Mutual Information Maximization ([MIM](#))), tested separately. In all the four approaches, GITA and Neurovoz² were analyzed separately, each database containing three repetitions of the sustained vowel /a:/ per speaker.

Regarding the front-end, all the voice signals were normalized and windowed using a Hamming window. The classification technique in this experimental set was [GMM](#), with G varying in powers of 2 within the range $\{4, \dots, 256\}$. A k -folds cross validation scheme was followed with $k=11$. All the features from the training folds were normalized in the range $[0, 1]$ and the same normalization parameters were employed for the testing folds. The threshold for the scoring was obtained using the [EER](#) point from the training folds.

Finally, a combination between the approach leading to the best results and the allophonic distillation approach was carried out.

10.2.1 First approach: noise, jitter, shimmer and complexity

In the first approach, noise, jitter, shimmer and complexity measures are used as input features of the classification system.

The noise features quantify the additive noise present in the phonatory signal which can be caused by an incomplete closure of the vocal folds and uneven movements of these, directly related with the rigidity in the muscles involved in phonation, caused by [PD](#). The employed features were [NNE](#), [HNR](#) and [GNE](#), calculated as it is indicated in [\[237\]](#), [\[238\]](#) and [\[239\]](#).

Two frequency and amplitude perturbation measurements were used too: jitter and shimmer. These allowed to quantify the irregular vibration of the vocal folds regarding the cycle-to-cycle variations of frequency and amplitude of the voice signal during phonation, which can be a reflection of non-symmetric movements or other impairments in the movements of the folds such as the tremor, as a consequence, again, of rigidity, descoordinations or tremor in the larynx muscles due to [PD](#). As jitter feature, the jitter relative to the average fundamental frequency, $\%jitt$ was used. In the same sense, as shimmer feature,

² CzechPD could not be employed in this experimental set as the corpus does not include sustained phonations

Figure 10.6: Diagram of first approach in phonatory experimental set.

%shim, the shimmer relative to the average amplitude of the signal was selected.

Complexity features are usually employed to measure and characterize non-linearity behaviors in the vocal folds as well as the dynamics of the phonatory system by analyzing the voice signal. The used features are D_2 , Largest Lyapunov Exponent (LLE) [247], Hurst Exponent (HE) [248], [249], DFA [248], [249], RPDE [248], Approximate Entropy (ApEn) [250], Sample Entropy (SaEn) [251], modified Sample Entropy (mSaEn) [252], Gaussian Kernel Entropy (GKE) [253], Fuzzy Entropy (FuzzyEn) [254] and Permutation Entropy (PE) [255].

To compute all of these features, the sampling frequency of the signals was maintained, and for noise and complexity features, the frame length was 55 ms as in [256] and an overlapping of 50%. It is important to clarify that to compute jitter and shimmer the signal is not enframed. All the cycle peak points are calculated for the whole signal and the local values of *%jitt* and *%shim* are calculated for the correspondent 55 ms frames with the mentioned overlapping.

An outline of this round of trials is depicted in Figure 10.6.

10.2.2 Second approach: Modulation Spectrum features

In the second approach, the proposed MS characterization was utilized in the front-end of the PD detector in order to characterize the voice signal. The vectors of features per frame include MSH, MSW, LMR, CIL, RALA, MSP25, MSP75 and MSP95.

Figure 10.7: Diagram of second approach in the phonatory experimental set.

To compute all of the **MS** features, signals are resampled to 25 kHz (after filtering and downsampling the original signal), the frame length was 180 ms as in [243] and an overlapping of 50%. The optimum configuration detailed in [243] regarding the calculation of **MS** is employed in this experimental set too. This configuration consists on the use of the acoustic and modulation frequency ranges detailed in Table 10.1.

Table 10.1: Optimum configuration to calculate MS features, as detailed in [243]. All values are in Hz.

Feature	Maximum modulation frequency	Minimum acoustic frequency	Maximum acoustic frequency
MSH	80	200	9000
MSW	80	200	9000
CIL	80	0	2000
RALA	200	800	6000
MSP ₂₅	80	0	1800
MSP ₇₅	200	0	2000
MSP ₉₅	200	0	9000

An outline of this round of trials is depicted in Figure 10.7.

Figure 10.8: Diagram of third approach in the phonatory experimental set.

10.2.3 Third approach: Rasta-PLP

Considering that Rasta-PLP coefficients provide AUC values over 0.80 in experimental set 1, these features are considered here too.

To compute the coefficients, the sampling frequency of the signals was 16 kHz (after filtering and downsampling the original signal), the frame length was 10ms and an overlapping of 50%. The values used for N were 10 and 12. This configuration is the one providing the best results in the baseline employing a sustained vowel in experimental set 1, as it can be observed in Table 6.4.

An outline of this round of trials is depicted in Figure 10.8.

10.2.4 Fourth approach: all features

In the fourth approach, all of the features included in the first and the second approach are appended in order to create a single feature vector per frame. As in the first approach, noise and complexity features were computed with voice signals sampled at 44,1 kHz, frame lengths of 55ms as in [256]. Jitter and shimmer were calculated for the same periods, as specified in section 10.2.1. MS features were calculated as in the second approach, in which the voice signals were filtered and downsampled to 25 kHz, and frame length was 180ms, since this set-up provide the best results in [243]. The Rasta-PLP+ Δ + $\Delta\Delta$ features were computed using 10 ms frames and 12 coefficients. In all cases, the frame shift was 27.5 ms. In this way, the number of frames of each type of parameterization at a certain utterance, was the same and the different vectors could be appended. In this round of trials, before the

Figure 10.9: Diagram of the fourth approach in the phonatory experimental set.

classification stage, two techniques were utilized separately to reduce the feature vectors dimension. The first one consisted in a dimensionality reduction technique based on [PCA](#). Three different feature vector lengths were analyzed: 6, 10 or 14. The [PCA](#) models were generated for each fold using only the training data and were applied to the testing data. The second one was a feature selection technique: [MIM](#) [257]. The number of features obtained after applying this technique were 6, 10 and 14.

An outline of this approach is depicted in [Figure 10.9](#).

10.2.4.1 Fifth approach: combination with an articulatory approach

To end, the scores from the approach providing the best results, were fused with the scores of the articulatory scheme providing the best results of this thesis (experimental set 3, allophonic distillation) employing a logistic regression scheme. The speech task for the articulatory scheme was fixed text, the allophonic distillation type was plosive for GITA and fricative for Neurovoz. The allophonic distillation was applied only to the [UBM](#) database (Albayzin). This combination of experiments was aimed to take advantage of the complementarity of the phonatory and articulatory aspects in the detection of [PD](#).

10.3 RESULTS

Tables 10.2 and 10.3 include the results of the four approaches for GITA and Neurovoz respectively. In all cases, the p-value associated to the correlations between the scores and the rating scales UPDRS and H-Y were over 0.05. The best results when employing PCA dimensionality reduction were obtained with feature vectors of dimension 14, after the reduction. Table 10.4 includes the results of the combination of the phonatory approach combining all the features (fourth approach) with PCA and feature vectors of dimension 14, after the dimensionality reduction. Figure 10.10 shows the ROCCH-DET curves of this combination. Best results are shaded in green.

Table 10.2: Best results in the four approaches for GITA database.

Approach	Accuracy $\pm$ CI	AUC	Sens.	Spec.	G	UPDRS ρ	H-Y ρ
Noise, %Jitt, %Shimm, complexity	74 $\pm$ 9	0.76	0.8	0.68	4	0.12	0
MS	69 $\pm$ 9	0.75	0.74	0.64	4	-0.19	-0.17
Rasta-PLP+ Δ + $\Delta\Delta$	72 $\pm$ 9	0.81	0.72	0.72	8	0.02	0
All features+PCA	71 $\pm$ 9	0.8	0.72	0.7	16	-0.03	-0.02
All features+MIM	74 $\pm$ 9	0.75	0.76	0.72	16	0	-0.02

Table 10.3: Best results in the four approaches for Neurovoz database.

Approach	Accuracy $\pm$ CI	AUC	Sens.	Spec.	G	UPDRS ρ	H-Y ρ
Noise, %Jitt, %Shimm, complexity	63 $\pm$ 11	0.61	0.67	0.58	4	0	-0.12
MS	59 $\pm$ 11	0.66	0.69	0.45	4	0.12	-0.06
Rasta-PLP+ Δ + $\Delta\Delta$	64 $\pm$ 10	0.63	0.75	0.48	4	0.18	-0.1
All features+PCA	64 $\pm$ 10	0.68	0.75	0.48	16	0.1	-0.17
All features+MIM	60 $\pm$ 11	0.64	0.63	0.58	4	-0.25	-0.15

Table 10.4: Results of combination of phonatory and articulatory approaches in GITA and Neurovoz. (*) the p-value associated to these correlation values was equal or below 0.05

Database	Accuracy $\pm$ CI	AUC	Sens.	Spec.	UPDRS ρ	H-Y ρ
GITA	85 $\pm$ 7	0.91	0.82	0.88	0.08	0.1
Neurovoz	87 $\pm$ 7	0.94	0.85	0.91	0.38*	0.27

Figure 10.10: ROCCH-DET curves of the combination of articulatory and phonatory approaches.

10.4 DISCUSSION

In this experimental set, several approaches to automatically detect PD using a sustained vowel were analyzed. Different features used in literature, such as noise, jitter, shimmer and complexity measurements along with PLP coefficients were analyzed in a GMM classification scheme. Other new features obtained from the MS are proposed and analyzed similarly. These new proposed features obtained better AUC than the rest of the features in the Neurovoz database (0.66 vs 0.61 and 0.63), however these results are not reproduced in the GITA database (0.75 vs 0.76 and 0.81). The combination of noise, jitter, shimmer and complexity provides the best accuracy results in GITA although Rasta-PLP and the combination of all features provide better AUC. In general, results are dissimilar in both databases and only the combination of all features with PCA dimensionality reduction is within the better results in terms of accuracy and AUC in both databases. This analysis suggest that none of these measurements can provide a high accuracy and reliability in the detection of PD on spite of the fact that, as it was explained at the beginning of this chapter, many patients present signs of dysphonia caused by PD.

The results obtained in this experimental set suggest that this type of measurements are complementary or can support other more complex measurements or approaches since the approaches analyzed in this chapter provide less differentiation properties to be used during diagnosis. This conclusion is supported by the fact that some phonatory perturbations or dysphonia are found too in elder neurologically healthy subjects. In the same sense, a signal with a relatively simple structure as a sustained vowel, contains much less information about motor problems than other more complex signals such as running speech. Therefore, it is coherent that the accuracy and AUC results of

this experimental set are usually below those obtained in the articulatory approaches presented in the previous chapters.

The combination of the articulatory and phonatory approaches provide a small increment of the *AUC* in the Neurovoz database (in the articulatory approach, this value was 0.93 and in the combined approach it is 0.94) although the accuracy, in this case is reduced in two absolute points. The results in the GITA database after the combination remain with the same values obtained using only the articulatory approach.

10.5 SUMMARY

In this experimental set, several automatic detectors of *PD* based on phonatory aspects of the speakers are analyzed. All of them employ a *GMM* classification scheme with different types of features, from which those extracted from *MS* are proposed in this thesis.

Results suggest that a combination of all the analyzed features with a *PCA* dimensionality reduction and a *GMM* classification scheme produce the best results in terms of accuracy, *AUC* and sensibility. Also that these type of approaches cannot provide results as higher as those obtained with the articulatory approaches proposed in this thesis.

Cross-validation trials (k-folds) employing exclusively phonatory information provide accuracies between 64% and 71%, with *AUC* between 0.68 and 0.80 depending on the database.

Part IV

DISCUSSION, FUTURE WORK AND FINAL CONCLUSIONS

DISCUSSION

In this chapter, an analysis and discussion of all the results obtained in the five experimental sets is carried out. Initially a summary of the experiments is included in section 11.1, followed by a table with the best results of each experiment and a comparative of ROCCH-DET curves, that are presented in section 11.2. Finally, the analysis of results and the global discussion is included in section 11.3.

11.1 EXPERIMENTAL SETS' SUMMARY

In this work, five main experimental sets destined to support the differential evaluation of PD by means of voice and speech processing were analyzed, each one containing different types of approaches, as it is listed in Table 11.1. The first four sets were mainly focused in the analysis of the articulatory aspects of speech to automatically detect PD by using speaker recognition techniques (Exp. 1), forced Gaussian mixtures at phoneme level (Exp. 2), and Gaussian models obtained after the selection of specific frames by performing allophonic distillation (Exp. 3) or acoustic landmark distillation (Exp. 4), employing in all cases different types of speech tasks. Finally, the last experimental set analyzed the use of phonatory aspects to detect PD employing only sustained vowels (Exp. 5).

Table 11.1: List of experimental sets. Exp. set stands for experimental set and Distill. stands for distillation.

Exp. set #	Name	Brief description	Approaches
1	Speaker recognition technologies	Use of state-of-the-art speaker recognition technologies: GMM-UBM and <i>i-Vectors</i> with different types of features.	1. fGMM
2	Forced Gaussians	Use of forced alignment techniques to segment speech and create separate Gaussian mixtures for each phonetic unit.	2. GMM-fUBM 3. fGMM-fUBM 4. opt-fGMM-fUBM
3	Allophonic distillation	Segmentation of speech by means of forced alignment techniques to perform frame selection attending to the type of allophone: fricative, liquid, nasal, plosive or vowel. The frames are employed in different GMM-UBM -based classification schemes.	1. Distill. in adaptation database 2. Distill. in adaptation and UBM database 3. Distill. in UBM database
4	Acoustic landmarks	Segmentation of speech by means of acoustic landmark detection to perform frame selection attending to the type of transition between allophones: burst, sonorant or glottal. The frames are employed in different GMM-UBM -based classification schemes.	1. Distill. in adaptation database 2. Distill. in adaptation and UBM database 3. Distill. in UBM database
5	Phonatory features	Use of phonatory features such as jitter, shimmer, noise, complexity, PLP and Modulation Spectrum features. Different GMM -based classification schemes are analyzed.	1. Jitter, shimmer... 2. MS 3. Rasta- PLP 4. All features

11.2 BEST RESULTS OVERVIEW

This section presents an overview of the results of the five experimental sets in order to facilitate the analysis and comparison of all the proposed approaches. Thus, Table 11.2 summarizes the best results of the five experimental sets (k-folds cross-validation) where the trials providing the best joint accuracy, [AUC](#) and sensibility per database and speech tasks are shaded in green and best global results per database are in bold. Table 11.3 contains the best results for cross-database trials, where best results are shaded in green. Figure 11.1 includes the [ROCCH-DET](#) curves for the four experimental sets analyzing the articulatory aspects in the three parkinsonian databases (k-folds cross-validation trials). Figure 11.2 includes the curves from the cross-database trials in which two parkinsonian databases are employed to adapt the [UBM](#) and

the third one, to test the resulting model. In these figures, the curves of Experimental set 5 are not incorporated since a different acoustic material was employed in this set and since Table 11.2 shows clear differences in performance between the phonatory approach and the rest of proposed approaches.

Figure 11.1: ROCCH-DET curves for the three parkinsonian speech databases employing the best set-ups for each experiment set. Fixed texts are selected for GITA and Neurovoz whereas DDK task from CzechPD is used.

11.3 DISCUSSION

An analysis of the tables 11.2 and 11.3 along with their correspondent ROCCH-DET curves helps revealing which speech task and approach can be more adequate to support the differential diagnosis of PD by means of speech and voice processing. Although each chapter referred to a specific experimental set contains a detailed discussion, in this chapter, a general discussion comparing all of them is introduced.

Firstly, it is important to remark that in most of the results of the trials carried out using GITA and Neurovoz, the CI is over 7%, pro-

Table 11.2: Best results for each experiment as a function of the employed database and speech task. Sust. v. stands for sustained vowel, and Exp. for Experimental set. * indicates p-Values equal or under 0.05.

Database	Speech task	Exp.	Accuracy $\pm$ CI	AUC	Sens.	Spec.	UPDRS ρ	H-C ρ
GITA	Fixed text	1	80 ± 8	0.85	0.82	0.78	0.18	0.25
		2	81 ± 8	0.88	0.84	0.78	0.32*	0.22
		3	85 ± 7	0.91	0.82	0.88	0.13	0.08
		4	85 ± 7	0.89	0.84	0.86	0.13	0.11
	DDK	1	81 ± 8	0.88	0.82	0.8	0.07	0.03
		2	79 ± 8	0.86	0.86	0.72	0	0
		3	83 ± 7	0.89	0.86	0.8	-0.04	-0.07
		4	83 ± 7	0.88	0.86	0.8	-0.01	-0.06
	Monologue	1	80 ± 8	0.88	0.76	0.84	0.34*	0.16
		2	78 ± 8	0.84	0.73	0.82	0.38*	0.25
		3	82 ± 8	0.89	0.8	0.84	0.42*	0.28*
		4	80 ± 8	0.86	0.76	0.84	0.38*	0.23
	Sust. v.	5	71 ± 9	0.8	0.72	0.7	-0.03	-0.02
	Fixed text + Sust. v.	5	85 ± 7	0.91	0.82	0.88	0.08	0.1
Neurovoz	Fixed text	1	86 ± 8	0.93	0.87	0.84	0.39*	0.12
		2	81 ± 9	0.87	0.83	0.78	0.3	0
		3	89 ± 7	0.93	0.87	0.91	0.4*	0.19
		4	89 ± 7	0.93	0.91	0.84	0.31	0.05
	DDK	1	79 ± 9	0.85	0.87	0.65	0.22	0.07
		2	79 ± 8	0.86	0.86	0.72	-0.11	-0.01
		3	86 ± 8	0.88	0.89	0.81	0.19	0.07
		4	83 ± 7	0.88	0.86	0.8	-0.01	-0.06
	Monologue	1	79 ± 12	0.81	0.59	0.9	0.06	0.06
		2	66 ± 14	0.67	0.35	0.83	0.17	0.06
		3	77 ± 12	0.79	0.53	0.9	0.45	0.2
		4	79 ± 12	0.9	0.47	0.97	0.33	0.3
	Sust. v.	5	64 ± 10	0.68	0.75	0.48	0.1	-0.17
	Fixed text + Sust. v.	5	87 ± 7	0.94	0.85	0.91	0.38	0.27
CzechPD	DDK	1	88 ± 1	0.94	0.85	0.93	0.02	0.01
		2	94 ± 1	0.97	0.9	1	0.04	0.01
		3	94 ± 1	0.98	0.9	1	-0.06	-0.19
		4	94 ± 1	0.99	0.9	1	0.27	0.12

Table 11.3: Best cross-database results for all the experiments. In all cases, speech task is DDK.

Test database	Exp.	Accuracy $\pm$ CI	AUC	Sens.	Spec.	UPDRS corr.	H-C corr.
GITA	1	73 $\pm$ 9	0.82	0.84	0.62	0.02	-0.22
	2	66 $\pm$ 9	0.76	0.9	0.42	0.22	0.11
	3	75 $\pm$ 8	0.84	0.86	0.64	-0.05	-0.13
	4	73 $\pm$ 9	0.81	0.8	0.66	0.12	0.04
Neurovoz	1	75 $\pm$ 10	0.82	0.8	0.65	0.33	0.33
	2	74 $\pm$ 10	0.78	0.87	0.5	0.2	0.09
	3	81 $\pm$ 9	0.83	0.91	0.62	0.32	0.22
	4	76 $\pm$ 10	0.83	0.78	0.73	0.46	0.31
CzechPD	1	79 $\pm$ 14	0.91	1	0.5	0.21	-0.01
	2	76 $\pm$ 14	0.87	0.95	0.5	0.12	0.07
	3	82 $\pm$ 13	0.95	0.85	0.79	0.21	-0.03
	4	82 $\pm$ 13	0.95	1	0.57	0.25	0.17

ducing overlapping of the results margins, reducing the certainty of some conclusions. Therefore, the analysis of the results included in this chapter is based on the observed tendencies rather than in precise and separated values and therefore do not allow to extract definitive conclusions.

In the same manner, in this discussion and in the discussions of each chapter, the best results are considered those in which the accuracy, AUC and sensitivity are higher. Sensitivity has been considered important too since, in an hypothetical clinical environment, the false rejections will be more undesirable than the false acceptances. In all of the cases, these values have been obtained for the EER operating point although the results will change for other points, depending on the requirements of clinical scenarios in which sensitivity might be increased.

11.3.1 The glass ceiling in the automatic detection of PD from speech

Although the minimum error found in all the analysed detectors in this thesis ranges from 6% to 15%, depending on the database under study, the theoretical limit of false rejection for these types of works is not clearly delimited and does not have to be 0% necessarily, like in other applications such as speaker recognition. Some specialists suggest that a 90% of PD patients suffer from dysarthria after a median latency

Figure 11.2: ROCCH-DET curves for the three parkinsonian speech databases used for testing in the cross-database trials, employing the best set-ups for each experiment set. In all cases, DDK tasks from the parkinsonian databases are used.

period of 7 years since diagnosis [41], [47]. These considerations would limit the false rejection in automatic detectors employing speech to 10%. Recent works like [51] have studied the presence of signs in the voice of PD patients and have quantified the percentage of affected patients to 100% using the Robertson dysarthria profile [52] in a cohort of 48 patients in several stages. However, these works are focused in perceptual estimations or preliminary quality of voice analysis. None of them have studied in detail the perturbations in voice and speech during the early stages neither the differences between the dysarthric disturbances caused by PD and by ageing. Thus, it is not possible to compare results with the minimum theoretical limit for false rejection but this value depends on the stage in which the patients included in the database are. This limit could be higher than 0% as the percentage of patients without perceived dysarthria (most of them in early stages) but with true deviations in articulation is unknown.

In the same manner, some studies point out that between 2% and 4% of patients diagnosed with idiopathic PD could be really suffering from other type of pathology, as autopsies reveal [18], [194]. Also, these values are obtained after the death of the patient and, since the diagnosis is more precise as the disease progresses, these error rates will be higher for populations with patients in the early stages. This means that some of the patients included in a parkinsonian speech database could not be really affected by PD, being this more likely in the newly diagnosed ones.

These two considerations would impose a minimum false rejection percentage in automatic detectors to values which could be close to 10%.

11.3.2 Features and kinetic changes

The analysis of results of Experimental set 1, employing classification techniques such as GMM-UBM or *i-Vectors*, suggests that PLP coefficients provide better results in the detection of PD than other parameterization features tested in that set. In fact, other features such as Frequency Domain Linear Prediction (FDLP) [258], Instantaneous Frequencies Cepstral Coefficients (IFCC) [259] or the acoustic cues from the acoustic landmarks were analyzed in preliminary trials during the thesis period and, in all cases, PLP outperformed the results provided by the other features¹. The use of Rasta-PLP and their associated derivatives, $\Delta + \Delta\Delta$, allows to characterize different types of dimensions. Firstly, derivatives provide information about velocity and acceleration of articulators, yielding cues about the presence of the disease, as literature has already demonstrated. Secondly, Rasta-PLP coefficients characterize consonants and this information can be used to detect the presence of the disease as a considerable portion of patients have problems while articulating consonants, specially those requiring more constriction of articulators as plosives and fricatives. Lastly, Rasta-PLP characterize formants in vowel segments and literature reveals that formants can be used to detect the influence of PD on speech as those indirectly indicate the position of the tongue during articulation. Although most of the works use VSA or similar features with these purposes, the use of features as PLP characterizing the formants and combined with machine learning techniques such as GMM can yield similar results.

The optimal analysis frame length is between 10 and 15 ms, according to this experimental set. Regarding the kinetic changes, analyzed in the Experimental set 1, results suggest that the recommended derivative segment length is between 40 and 45 ms. In this sense, results of Experimental set 1 suggest that the use of derivatives is crucial in the detection with the proposed approaches since in the trials in which

¹ These preliminary trials were not included in the thesis document for the sake of conciseness.

these are not used along with the Rasta-PLP coefficients, the accuracy is reduced in a 17% when using the GMM-UBM classification scheme.

11.3.3 Analysis of experimental sets

The first experimental set is considered the baseline of this thesis and provides accuracy results between 81% and 91%, with AUC between 0.88 and 0.94 depending on the employed database. In general, as it can be observed in the ROCCH-DET curves of Figures 11.1 and 11.2, this approach provides accuracies over 79% at EER operating point in the detection of PD but not the best among all the performed tests.

The techniques proposed in the Experimental set 2 can be considered novel, specially for the detection of PD, although previous works such as [260] used a similar approach for speaker recognition employing DNN posteriors instead of labels from speech forced alignment to create the UBM in an *i-Vectors* classification scheme. The results obtained in this experimental set are in general lower than those provided by the rest of the articulatory approaches, indicating that these techniques are not the most advisable for the automatic detection of PD through speech. Nevertheless, the main advantage of this set is that it allows to observe the influence of PD in each of the individual phonetic units. The results of this experimental set suggest that, in general, phonetic units requiring a higher narrowing of the vocal tract but without a burst tend to be more influential in the detection of PD while nasal consonants are the less influential. However, comparing the results obtained with Neurovoz and GITA, these suggest that PD affects to the articulatory sequence as a whole.

From all the articulatory-based approaches, two new techniques and their respective denominations, are proposed in this thesis: allophonic distillation (Experimental set 3) and acoustic landmark distillation (Experimental set 4). While both techniques are based on frame selection, the first one aims to the use of specific segments of the speech related to the manner of articulation, such as plosives or fricatives, while the second one is focused in the selection of certain types of transitions (from silence to plosive or from vowel to consonants, among others).

The use of allophonic or landmark distillation in the UBM produces improvements respect to the baseline (Experimental set 1) in the three databases and, practically, with any speech task. Attending to the results shown in Tables 11.2 and 11.2 the results of Experiment 3 (allophonic distillation) tend to outperform the rest of the approaches in the k-folds cross-validations and in the cross-database validations, for all databases and speech materials, except in the case of Neurovoz database and monologues in which the acoustic landmark distillation (Exoerimental set 4) provides better accuracy and AUC. This approach can lead to supervised schemes when using DDK tasks and fixed texts or to unsupervised schemes when using monologues to train and test the systems. In Experimental sets 3 and 4, the accuracy and AUC were,

in general, higher only when applying the distillation in the [UBM](#). This behavior is reproduced in the two experiments with all the speech tasks. Thus, the distillation in the adaptation-testing databases (leading to supervised schemes) serves to identify which segments or transitions are the most influential in the detection of [PD](#) through speech but do not always lead to the highest accuracy. The reason why the most successful approaches are those in which the distillation is applied to the [UBM](#) is substantiated in the fact that when, for instance, distilling the [UBM](#) database to contain only fricative segments, the resulting [GMM-UBM](#) tends to model the fricatives with more detail but does not discard other information contained in the speech from the adaptation-testing database (parkinsonian databases) that complements the fricative segments with others containing information as articulation of vowels, plosives, formants, amplitude envelope, and other characteristics represented directly or indirectly by the Rasta-[PLP](#) coefficients and its derivatives.

The results of Experimental set 2 show that performing a comparison of patients and controls at a phonetic unity level can provide a high accuracy. However, the approaches from Experimental sets 3 and 4 using more general groups of segments, like fricatives or b-lmk transitions, yield better accuracies and [AUC](#). In a similar way, on Experimental set 2 it is unclear which phonetic unity provides the best results as there is not clear correspondence in the results obtained with GITA and Neurovoz. Only the labels /R/ and /D/ tend to produce better sensitivity and specificity. However, in Experimental set 3 the influence of plosives in the detection of [PD](#) is more clear in GITA and Neurovoz simultaneously.

The [ROCCH-acDET](#) curves depicted in Figures [11.1](#) and [11.2](#) show that the performance of Experimental sets 3 and 4 are similar in most of the experiments and are, almost always, over the performance of Experimental sets 1 and 2. In some of the trials, Experimental set 3 provides higher accuracy for lower miss probability (higher sensibility) while in other cases this occurs with Experimental set 4.

Regarding the fusion of scores employing logistic regression, in Experimental sets 3 and 4, fusing does not produce the same effects in all the databases or speech tasks and, in some cases, it reduces the accuracy or [AUC](#).

Concerning the cross-database tests, which have been carried out using only [DDK](#) tasks as is the only task which can be used with CzechPD employing Albayzin as [UBM](#), as it is observed in Table [11.3](#) the reductions of the best [AUC](#) respect to the k-folds cross-database trials never exceed 0.05 absolute points. Also, the use of cross-database trials for the automatic detection of [PD](#) is a novelty by itself since, to the author of this thesis knowledge, no published work has performed these type of trials.

Respecting the phonatory approaches of Experimental set 5, it is remarkable the fact that in some examples found in literature, such as

[99], values over 98% of accuracy are obtained using sustained vowels and, mainly, MFCC in a SVM classifier while in this thesis, using the same type of features (in Experiment 1) and a GMM-UBM classifier, the maximum accuracy does not exceed 76% values. Something similar occurs with work [103] where authors report a 91% of accuracy employing MFCC coefficients and a SVM classifier. These differences between the results of this thesis and [99] could be caused not only by the use of different databases but by the fact that in the cited work, authors use the voices of the same speakers in the training and test sets and this could lead to over-optimistic results. Regarding [103], the differences can be due to the use of different databases (in the cited work only one sustained vowels from 34 speakers, 17 patients) and classification schemes. In any case, most of the trials in [103] provide results under 80% and only an isolated trial provides a 91% accuracy. The results of Experimental set 5 coincide with those found in [101] where a 74% accuracy is obtained using a sustained vowel /a/ at comfortable SPL. However, in that work more vowels and phonatory levels are employed, reaching 90% accuracy, suggesting that the phonatory approaches proposed in this thesis could provide better results in other scenarios.

Tables B.1, B.2 and B.1 in Appendix B show the global scores of each speaker in the testing folds of the trials leading to the best results on all the experimental sets and databases. These show the classification of the speakers as parkinsonian (cells shaded in red) or control (cells shaded in green) and allow to confirm the analysis exposed previously.

Globally, the highest accuracy obtained in the trials is 94% with AUC of 0.99, sensitivity of 0.90 and specificity of 1.00, in the CzechPD database using acoustic landmark allophonic distillation on the UBM (third approach) presented in the Experiment 4 with fusion of scores (b-lmk with -s-lmk). It is important to consider that this database mainly contains newly diagnosed patients. Therefore, results suggest that the new proposed approaches can be valid detecting PD in early stages. The best results obtained in the k-folds cross-validation in the three databases are over 85%, however, the lowest accuracy and AUC among the optimum results are occur with GITA, the database with the largest number of speakers and with the best balance in age and sex.

With this analysis it is possible to determine that the most advisable approaches to support the differential diagnosis of PD, in terms of accuracy, AUC and sensibility, are those employing allophonic or acoustic landmark distillation in the UBM within GMM-UBM classification schemes, as described in Experimental sets 3 and 4. However, it is unclear which one of these two provide better results.

11.3.4 Speech tasks

Equally, from all the speech tasks, fixed texts provide the best results in almost every trial, as it can be inferred from Table 11.2. This can be ex-

plained in the fact that, as observed in the results of Experimental sets 2, 3 and 4, the impact of PD on speech is not limited to a few articulatory movements, allophones or transitions but influences them all in a higher or lower degree. Fixed texts contain more variety of allophones and articulatory movements than DDK tasks and that justifies the differences in the results for both tasks. On the other hand, although monologues include a larger variety of articulatory movements too, these are different for each utterance in the training and testing sets and the resulting models are text-independent. And in this case, text-dependent models demonstrate to outperform to text-independent approaches as the first ones allow to compare more precisely the specific segments (transitions, allophones, acceleration or velocity of articulatory movements, etc).

Monologues tend to provide much less sensitivity than specificity in all the experimental sets, as it is observed in Table 11.2. However, the same table shows that these speech tasks can be more advisable to predict rating scales such as UPDRS. However, fixed texts and DDK tasks provide higher sensitivity than specificity in the vast majority of the trials suggesting that these materials are more adequate for the differential diagnosis of PD with the proposed approaches.

11.3.5 Scores and rating scales

Regarding the correlation between the global scores provided by the classifiers for each patient and their correlation with the UPDRS ratings, these values rarely exceed 0.5. However, this correlation is being calculated with the global rating while using only the UPDRS motor part would have been much more appropriate. However, only the global ratings were available for the three databases. Nevertheless, these results serve to observe that the higher correlations are obtained with monologues, suggesting that these speech tasks are more appropriate for systems evaluating the grade of affection of the disease. In general, correlations between scores and H-Y ratings are non-significant.

11.3.6 Other considerations

In Tables B.1, B.2 and B.1 in Appendix B it is possible to observe that the miss-classifications occur more often in patients with low UPDRS although, in some cases these are manifested in patients with high UPDRS diagnosed a long time ago. This suggest that the proposed approaches could be less sensitive with individuals in their initial stages but not necessarily.

Additionally, the signs of this disease can vary along the day depending on the individual and these can appear and disappear spontaneously. Hence, the use of several recordings of a subject, carried out at different moments or in different sessions will produce a more ro-

bust detection system. However, in this thesis all the recordings from a given speaker are recorded in the same session, increasing the uncertainty.

One of the main drawbacks of the analysis of the results included in this thesis, and therefore, of the possible conclusions that can be extracted from this analysis, is that these are limited to the three databases included in this thesis. In some cases, the use of some of the proposed methodologies produces the same improvements using three parkinsonian databases separately, suggesting that the methodology is valid for the desired purposes.

However, Neurovoz is unbalanced in number of speakers per class (there are more patients than controls) and more males, in total, being female patients 2.5 years older than female controls and male patients 5.3 years older than male controls. These facts, specially the one related with age, can bias the results since speech is influenced by age.

In general, it is important to consider that all of these experiments are analyzed employing only idiopathic **PD** patients and controls and no other parkinsonism or neurodegenerative disease such as ataxia or **HD** influencing speech has been considered. Consequently, it is not possible to evaluate if the proposed approaches provide discrimination between idiopathic **PD** and other neurodegenerative diseases.

In general, it is difficult to compare these results with others in the literature since the used databases, materials and methodologies are so diverse that the comparison of accuracies between the ones obtained in this work and the ones found in other studies will not lead to any conclusion. However, as it is said in the analysis of the state-of-the-art of this thesis, the most relevant and recent works provide an accuracy varying between 80% and 95%, coinciding with the results exhibited here.

CONTRIBUTIONS, FUTURE WORK AND CONCLUSIONS

In this last chapter, the contributions of the thesis are listed, followed by the future work and the conclusions derived from the analysis of the results included in the specific chapters, compared and combined in the discussion chapter.

12.1 CONTRIBUTIONS

The contributions of the work presented in this doctoral thesis are mainly related to the proposal of new, state-of-the-art methodologies to support the differential diagnosis of [PD](#) during the clinical assessment of patients based on articulatory aspects of speech, since these types of aspects are the most unexplored in the literature. These contributions are listed as follow:

1. A thorough analysis of state-of-the-art speaker recognition technologies, identifying the optimum set-up for the detection of [PD](#).
2. A new database of parkinsonian and control speakers has been recorded with the associated demographic and clinical metadata.
3. The proposal of a new scheme based on allophonic distillation, providing similar results in the analyzed parkinsonian databases.
4. An analysis of the role of the distinct allophonic types (fricative, liquid, nasal, plosive and vowels) in the automatic detection of [PD](#) in Spanish language. Identification of the most relevant segments.
5. An analysis of the role of the distinct phonetic units in the detection of [PD](#) in Spanish language. Identification of the behavior of these units in the proposed detection systems and their relation with the disease.
6. An analysis of the different types of transitions between segments during speech articulation and their influence in the automatic detection of [PD](#).
7. The analysis and identification of several new supervised and unsupervised schemes to detect [PD](#) from articulatory aspects of speech, depending on the used speech task and the front-end.
8. The identification of the advantages and disadvantages of the use of fixed texts, monologues, sustained vowels and [DDK](#) tasks within the proposed schemes to detect [PD](#).

9. The analysis of the combination of a phonatory subsystem with an articulatory subsystem to automatically detect [PD](#).
10. The study of the relevance articulators' kinematics in the automatic detection of [PD](#).
11. The study of schemes that under certain circumstances can be considered as language-independent systems to detect [PD](#) from speech.
12. The study of the generalization properties of the analyzed systems by applying cross-database validation schemes.

12.2 FUTURE WORK

Since in this thesis, phonatory and articulatory approaches are proposed and analyzed, new prosodic approaches must be assessed in the future, as prosody is the only motor speech aspect not explored directly in this thesis. To this respect, the use of phonetograms using different types of voice or speech tasks in parkinsonian speakers is almost unexplored. This tool, providing information about frequency and speech [SPL](#) ranges, which have demonstrated to be useful in the assessment of the disease, can be used in automatic detection systems and at the same time, can provide visual and easily interpretable information to clinicians, as it has been stated in [[141](#)].

As the distinction between the different parkinsonian syndrome variants is crucial to set an adequate long-term treatment for the patient, new objective indexes should provide a differentiation between the different parkinsonisms or other neurodegenerative diseases with similar symptoms, such as ataxia or [HD](#). In order to analyze the peculiarities and differences between the voices of patients of the distinct neurodegenerative diseases, new databases must be designed and recorded including a high number of already diagnosed patients of each neurodegenerative disease. In the same sense, from all the databases studied in the state-of-the-art chapter, only a few contain several speech tasks from English parkinsonian speakers. Therefore, new English speaker databases containing speech tasks according to the findings highlighted in this thesis must be developed.

On the other hand, Neurovoz must be complemented with more patients and controls in order to balance the two classes in terms of sex, age and number of speakers per class.

As age influences speech, new classification systems containing models for specific age ranges must be analyzed, in order to separate the effects of age and [PD](#) over the voice and speech of speakers.

The study of unsupervised systems using monologues as speech task with longer recordings, allowing to create more precise models of the parkinsonian and control classes must be assessed. These systems

will be of use for non-clinical environments related to the emerging speech assistant technologies.

Moreover, the phonatory approaches must be evaluated again employing more vowels at different phonatory levels.

The proposed methodologies in this thesis must be validated with other databases and in other languages to verify if these are language-independent or not. It is expected that some parameters such as the optimal τ_{frame} and $\tau_{derivative}$ will be slightly different in other languages as languages contain different allophones and, all of them contain their own articulatory particularities. Concretely, the adaption of the allophonic distillation techniques to other languages must be carried out in collaboration with other research groups or phoneticians to validate the methodologies involving phonetic knowledge distinct from Spanish, in order to avoid methodological errors regarding the grouping of phonemes.

Regarding the cross-database results, although best accuracy in this thesis reaches promising values, as it can be inferred from Table 11.3, compensation techniques to eliminate the effect of the channel must be included to produce more robust models. In the same sense, new score normalization techniques for two-class cases like the studied in this thesis must be explored.

In order to provide clinicians with tools to help in the diagnosis of PD, the inclusion of a-priori probabilities of a subject to suffer from PD as a function of their age, sex and other circumstances must be considered in a global approach.

The proposal of new biomarkers with a clearer physical interpretation is needed. These type of features are preferred in clinical environments since clinicians can extract conclusions derived from the analysis of the biomarkers in a more understandable and reasoned manner.

The analysis of the speakers voice and speech within a longitudinal study framework using the proposed systems must be carried out to validate the system, to estimate its consistency and to propose more robust solutions.

Additionally, after the analysis of the state of the art, it is clear that a large amount of works have analyzed the parkinsonian voice and speech since the second half of the 20th century. Despite of the evidences about the differences in phonatory, articulatory and prosodic features between patients and controls, the use of these methods has not been widely spread in the clinical practice with diagnosis purposes. One of the main problems is that there is not a single clear biomarker related to speech providing good and reliable results. Nevertheless, the results obtained in this thesis suggest that new machine learning techniques can help to overcome this issue by combining information coming from several aspects and features to generate a score. Therefore, a next step is the design of a protocol to obtain several speech-based features based on phonatory, articulatory and prosodic aspects to be

analyzed within a machine learning scheme to provide differential information to clinicians during the differential diagnosis. The main goal is to clinically test this system to prove its usefulness in true clinical environments as there are not published clinical trials of this type.

Lastly, the study of multimodal schemes combining speech with other biosignals as those extracted from the analysis of eye movements, gait or drawing, among others, must be addressed.

12.3 CONCLUSIONS

In this doctoral thesis, new approaches to support the differential evaluation of PD by means of voice and speech processing have been proposed and analyzed.

One of the first conclusions obtained in this thesis is that kinematic changes, characterized by means of the derivatives of PLP coefficients of speech are crucial in the detection of parkinsonian speech as these provide information about acceleration and velocity of the articulators during speech, two characteristics that are highly influenced by the disease in the forms of hypokinesia or other types of kinetic impairments such as akinesia or bradykinesia.

Since the motor dysfunctions caused by PD produce misarticulation, results suggest that these deficits can be found in different types of segments of speech (i.e., distinct groups of allophones and transitions). The articulatory movements requiring a higher narrowing of the vocal tract, that normally occur during the pronunciation of fricatives and plosives, are the most influenced by the disease. However, these are not the only segments of speech containing information about PD since schemes based only on vowels provide significant results too. Results suggest that PD influences all the studied phonetic units and groups, affecting to the articulatory sequence as a whole. For this reason, the analysis of phonetically balanced speech tasks allows to evaluate the presence of PD from speech by using automatic detectors. Additionally, when employing fixed texts as speech tasks, the obtained classification models are text-dependent and allow to compare more precisely the articulation of patients and controls since all the speakers repeat the same sequence of allophones. Consequently, phonetically balanced fixed texts are recommended for automatic detection systems in the clinical practice.

The discriminatory properties of the proposed phonatory approaches are quite limited in comparison with the articulatory approaches. The combination of these approaches yields better results in some cases but the advantages of this combination are unclear.

Higher correlations between scores of the classification systems and the UPDRS ratings are obtained with monologues, suggesting that these speech tasks are more appropriate for systems evaluating the grade of affection of the disease.

Results suggest that the optimum frame lengths range between 10 and 15 ms and the derivative segment length between 40 and 50 ms for approaches based on Rasta-PLP characterization and GMM-UBM classification schemes.

The proposed approaches exhibit a lower accuracy but a similar sensibility in the cross-validation trials with respect to the cross-validation (k-folds) trials, indicating that these approaches generalize in scenarios with different recording conditions.

The best accuracy results obtained with the proposed methodologies reach accuracy values ranging from 85% to 94% with AUC between 0.91 and 0.99 and sensitivity between 0.82 and 0.91 depending on the adaptation-testing database. These results are obtained employing the proposed allophonic distillation technique in the UBM database in a GMM-UBM classification scheme. These values can be considered close to the maximum feasible results which are under 100% accuracy. These true limits have not been considered before in other works performing automatic detection of PD.

Part V

APPENDIX

APPENDIX A

In this appendix, tables [A.1](#) and [A.2](#) containing the places of articulation, and number of joint vowels and consonants in Sentences 1 and 2 from GITA are included.

Table A.1: Places of articulation, and number of joint vowels and consonants in Sentence 1

[illegible]

APPENDIX B

In this appendix, the global scores obtained after the classification of the testing utterances are included in Tables [B.1](#), [B.2](#) and [B.3](#) for GITA, Neurovoz and CzechPD, respectively. Only the scores obtained within the trials leading to the best results at each experiment are included, in correspondence with Table [11.2](#). For experimental set 5, only the trials employing sustained vowels are considered. In the three tables, S. Code refers to the speaker code employed in all the experimental sets. If the code includes the suffix "PD", the speaker is a parkinsonian patient and if the suffix is "Ctl", the speaker is a control subject. In all the cases, the scores are shaded in red if it verifies the hypothesis of suffering from PD (the score is over the [EER](#) threshold) or in green, otherwise. All tables include [UPDRS](#) and [H-Y](#) scores as well as the number of years since diagnosis, referred as "Years".

Table B.1: Scores per speaker from the testing utterances in the approaches leading to best results of each experiment. Only fixed text from GITA database is considered.

S. Code	UPDRS	H-Y	Years	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	S. Code	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
A0001PD	28	2	3	0.31	0.39	0.47	0.51	0.45	C0001CH	0.64	0.34	0.29	0.3	0.25
A0002PD	19	1	2.5	0.42	0.6	0.58	0.56	0.48	C0003CH	0.17	0.45	0.43	0.44	0.42
A0003PD	52	3	3	0.91	0.96	0.84	0.81	0.8	C0004CH	0.54	0.48	0.4	0.45	0.34
A0005PD	32	2	12	0.71	1	0.83	0.86	0.65	C0005CH	0.15	0.2	0.21	0.16	0.11
A0006PD	28	2	4	0.61	0.46	0.43	0.54	0.49	C0006CH	0.56	0.34	0.33	0.37	0.33
A0007PD	30	2	12	0.55	0.79	0.68	0.79	0.54	C0007CH	0.32	0.37	0.17	0.18	0.23
A0008PD	29	2	7	0.55	0.47	0.38	0.39	0.34	C0008CH	0.38	0.12	0.38	0.4	0.2
A0009PD	41	3	37	0.69	0.73	0.69	0.75	0.69	C0010CH	0.043	0.24	0.16	0.17	0.19
A0010PD	38	3	41	0.41	0.63	0.48	0.6	0.55	C0011CH	0.33	0.43	0.49	0.48	0.78
A0011PD	43	3	12	0.55	0.23	0.32	0.33	0.13	C0012CH	0.28	0.45	0.27	0.27	0.14
A0013PD	6	1	8	0.54	0.29	0.38	0.41	0.32	C0013CH	0.15	0.39	0.4	0.47	0.28
A0014PD	61	3	17	0.59	0.77	0.78	0.81	0.44	C0014CH	0	0.25	0.31	0.35	0.27
A0015PD	28	2	4	0.57	0.49	0.58	0.61	0.33	C0015CH	0.28	0.35	0.38	0.43	0.17
A0016PD	44	3	10	0.56	0.64	0.73	0.83	0.37	C0016CH	0.41	0.51	0.45	0.54	0.27
A0017PD	50	2	12	0.27	0.8	0.9	0.93	0.34	C0017CH	0.48	0.35	0.43	0.48	0.32
A0020PD	30	2	12	0.57	0.81	0.65	0.73	0.5	C0018CH	0.31	0.25	0.066	0.15	0
A0021PD	42	3	12	0.4	0.53	0.49	0.51	0.46	C0019CH	0.043	0.17	0.12	0.19	0.39
A0022PD	20	2	0.4	0.41	0.21	0.53	0.52	0.25	C0020CH	0.36	0.15	0.074	0.099	0.25
A0023PD	14	1	1	0.7	0.96	0.48	0.54	0.38	C0021CH	0.32	0	0	0.045	0.19
A0024PD	92	5	15	1	0.96	0.95	1	1	C0022CH	0.33	0.27	0.33	0.43	0.26
A0025PD	53	2	7	0.28	0.4	0.53	0.58	0.57	C0023CH	0.3	0.12	0.28	0.34	0.28
A0026PD	21	1	4	0.42	0.59	0.6	0.73	0.41	C0024CH	0.26	0.38	0.29	0.36	0.17
A0027PD	13	1	1	0.36	0.39	0.33	0.34	0.37	C0025CH	0.2	0.37	0.43	0.49	0.36
A0029PD	19	2	12	0.5	0.61	0.43	0.44	0.24	C0026CH	0.2	0.35	0.34	0.44	0.2
A0030PD	75	3	16	0.62	0.81	0.74	0.72	0.47	C0027CH	0.34	0.4	0.41	0.42	0.64
A0031PD	40	2	14	0.5	0.4	0.47	0.54	0.39	C0028CH	0.43	0.76	0.45	0.49	0.26
A0032PD	30	2	14	0.66	0.51	0.69	0.68	0.55	C0029CH	0.28	0.26	0.15	0.18	0.24
A0034PD	19	2	17	0.57	0.51	0.55	0.55	0.27	C0030CH	0.26	0.31	0.3	0.33	0.14
A0035PD	23	2	10	0.43	0.56	0.54	0.5	0.26	C0031CH	0.59	0.25	0.41	0.44	0.13
A0037PD	40	2.5	12	0.35	0.57	0.58	0.64	0.44	C0033CH	0.46	0.25	0.3	0.36	0.33
A0038PD	9	2	12	0.29	0.24	0.43	0.52	0.16	C0034CH	0.29	0.11	0.22	0.31	0.26
A0039PD	67	4	20	0.62	0.7	1	0.98	0.61	C0035CH	0.012	0.19	0.24	0.33	0.38
A0041PD	15	3	4	0.48	0.57	0.66	0.67	0.23	C0037CH	0.32	0.55	0.63	0.74	0.34
A0042PD	54	3	8	0.57	0.48	0.55	0.63	0.3	C0039CH	0.31	0.6	0.58	0.7	0.31
A0043PD	51	2	9	0.56	0.7	0.82	0.82	0.3	C0040CH	0.24	0.25	0.27	0.39	0.34
A0045PD	71	3	17	0.48	0.56	0.59	0.67	0.51	C0041CH	0.027	0.33	0.49	0.52	0.17
A0046PD	40	3	5	0.6	0.79	0.78	0.86	0.54	C0042CH	0.026	0.22	0.38	0.43	0.37
A0047PD	40	2	3	0.77	0.62	0.64	0.66	0.14	C0043CH	0.3	0.14	0.21	0.31	0.37
A0048PD	53	2	16	0.49	0.45	0.52	0.46	0.23	C0044CH	0.32	0.3	0.37	0.48	0.25
A0049PD	28	2	4	0.52	0.87	0.73	0.75	0.37	C0045CH	0.36	0.2	0.16	0.19	0.092
A0050PD	38	2	4	0.42	0.45	0.46	0.44	0.2	C0046CH	0.38	0.28	0.29	0.33	0.22
A0051PD	57	2	1	0.41	0.42	0.47	0.5	0.47	C0047CH	0.59	0.31	0.29	0.3	0.28
A0052PD	23	2	12	0.8	0.57	0.71	0.77	0.54	C0048CH	0.16	0.32	0.31	0.36	0.26
A0053PD	33	2	2	0.26	0.71	0.64	0.62	0.41	C0049CH	0.38	0.42	0.36	0.45	0.35
A0054PD	30	2	7	0.031	0.22	0.055	0	0.052	C0050CH	0.068	0.17	0.28	0.38	0.033
A0055PD	53	2	19	0.58	0.62	0.72	0.7	0.54	C0051CH	0.53	0.49	0.59	0.72	0.2
A0056PD	45	2	3	0.46	0.36	0.64	0.61	0.41	C0052CH	0.48	0.26	0.56	0.55	0.21
A0057PD	65	3	8	0.64	0.54	0.49	0.55	0.4	C0053CH	0.18	0.28	0.22	0.26	0.33
A0058PD	21	1	7	0.24	0.12	0.23	0.24	0.36	C0054CH	0.28	0.39	0.22	0.31	0.22
A0059PD	29	2	43	0.68	0.44	0.55	0.56	0.087	C0057CH	0.29	0.25	0.38	0.4	0.14

Table B.2: Scores per speaker from the testing utterances in the approaches leading to best results of each experiment. Only fixed text from Neurovoz database is considered.

S. Code	UPDRS	H-Y	Years	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	S. Code	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
0004PD	59	4	9	1	1	1	1	0.58	0034Ch	0.78	0.93	0.81	0.71	0.53
0006PD	11	3	13	0.89	0.9	0.94	0.92	0.54	0036Ch	0.22	0.33	0.37	0.49	0.41
0007PD	13	2	2	0.63	0.8	0.72	0.7	0.53	0045Ch	0.36	0.33	0.35	0.4	0.56
0008PD	20	2	6	0.48	0.56	0.5	0.68	0.51	0048Ch	0.38	0.38	0.3	0.31	0.44
0009PD	13	2	4	0.7	0.82	0.68	0.78	0.53	0049Ch	0.22	0.24	0.27	0.28	0.0089
0010PD	24	2	1	0.64	0.65	0.72	0.66	0.54	0051Ch	0.41	0.51	0.45	0.58	0.5
0011PD	9	1	30	0.56	0.65	0.63	0.61	0.5	0052Ch	0.31	0.32	0.28	0.18	0.57
0012PD	9	2	6	0.8	0.72	0.77	0.79	0.45	0053Ch	0	0	0.037	0.23	0.45
0013PD	5	2	5	0.53	0.39	0.56	0.63	0.48	0054Ch	0.42	0.37	0.43	0.29	0.42
0014PD	31	2	8	0.75	0.68	0.8	0.72	0.8	0055Ch	0.51	0.5	0.38	0.56	0.41
0015PD	13	1	2	0.51	0.59	0.58	0.65	0.48	0056Ch	0.51	0.56	0.52	0.44	0.46
0016PD	-	2	-	0.78	0.84	0.83	0.76	0.56	0060Ch	0.31	0.35	0.37	0.27	0.49
0017PD	-	3	6	0.72	0.78	0.73	0.79	0.43	0061Ch	0.39	0.28	0.3	0.25	0.45
0018PD	12	2	3	0.55	0.54	0.55	0.61	0.49	0062Ch	0.11	0.093	0.087	0.21	0.42
0019PD	17	2	2	0.7	0.79	0.73	0.73	0.56	0063Ch	0.26	0.23	0.34	0.27	0.59
0020PD	-	-	10	0.79	0.79	0.78	0.88	0.33	0064Ch	0.46	0.48	0.45	0.36	0.44
0021PD	14	-	5	0.87	0.94	0.87	0.8	0.49	0065Ch	0.1	0.21	0.17	0.22	0.49
0022PD	-	2	11	0.85	0.78	0.88	0.8	0.6	0068Ch	0.21	0.12	0.13	0.098	0.4
0023PD	12	2	-	0.84	0.76	0.94	0.95	0.68	0069Ch	0.083	0.26	0.18	0.085	0.47
0024PD	19	2	-	0.79	0.78	0.82	0.82	0.51	0071Ch	0.59	0.71	0.62	0.48	0.48
0025PD	9	2	7	0.71	0.7	0.75	0.74	0.61	0072Ch	0.25	0.52	0.41	0.69	0.71
0026PD	26	3	-	0.8	0.8	0.96	0.73	0.59	0073Ch	0.47	0.6	0.56	0.37	0.63
0027PD	32	2	7	0.75	0.72	0.73	0.85	0.72	0074Ch	0.3	0.42	0.34	0.32	0.52
0028PD	10	3	8	0.49	0.29	0.43	0.68	0.36	0075Ch	0.022	0.15	0	0	0.34
0029PD	5	2	1	0.71	0.58	0.72	0.77	0.63	0076Ch	0.5	0.49	0.51	0.44	0.59
0030PD	9	2	10	0.63	0.83	0.78	0.84	0.45	0080Ch	0.47	0.64	0.54	0.44	0.59
0031PD	35	3	5	0.61	0.64	0.62	0.86	0.44	0081Ch	0.46	0.58	0.4	0.36	0.52
0032PD	22	2	9	0.67	0.76	0.8	0.72	0.62	0082Ch	0.096	0.28	0.18	0.098	0.65
0033PD	16	2	10	0.52	0.63	0.64	0.71	0.57	0083Ch	0.55	0.5	0.51	0.54	0.5
0035PD	14	2	10	0.74	0.82	0.87	0.66	0.48	0085Ch	0.41	0.34	0.33	0.24	0.18
0037PD	47	5	8	0.64	0.5	0.75	0.61	0.57	0086Ch	0.25	0.21	0.21	0.15	0.44
0038PD	10	2	-	0.59	0.47	0.62	0.64	0.5	0087Ch	0.47	0.38	0.45	0.37	0.33
0039PD	4	2	8	0.49	0.54	0.57	0.58	0.95						
0040PD	-	-	2	0.36	0.28	0.35	0.17	0.49						
0041PD	23	2	4	0.63	0.55	0.65	0.63	0.71						
0042PD	7	2	5	0.3	0.33	0.24	0.25	0.59						
0043PD	11	2	8	0.38	0.33	0.36	0.39	0.54						
0044PD	4	1	1	0.51	0.56	0.56	0.6	0.55						
0046PD	20	3	1	0.5	0.41	0.45	0.51	0.14						
0047PD	17	3	8	0.44	0.63	0.51	0.5	0.51						
0058PD	31	2	2	0.63	0.58	0.6	0.63	0.5						
0066PD	21	2	5	0.82	0.65	0.67	0.5	0.6						
0067PD	26	3	23	0.85	0.67	0.77	0.52	1						
0070PD	17	3	2	0.66	0.54	0.62	0.51	0.06						
0077PD	20	2	5	0.7	0.53	0.71	0.54	0.48						
0078PD	18	3	14	0.68	0.46	0.68	0.5	0						

Table B.3: Scores per speaker from the testing utterances in the approaches leading to best results of each experiment. Only [DDK](#) task from Czech database is considered.

S. Code	UPDRS	H-Y	Years	Exp. 1	Exp. 2	Exp. 3	Exp. 4	S. Code	Exp. 1	Exp. 2	Exp. 3	Exp. 4
01PD	20	2	2	0.82	0.8	0.83	0.8	23Ctl	0.44	0.25	0.35	0.3
02PD	32	3	4	0.7	0.72	0.67	0.65	24Ctl	0.46	0.26	0.31	0.26
03PD	18	2	1	0.94	0.94	0.9	0.94	25Ctl	0.49	0.29	0.46	0.29
04PD	12	2	1.3	0.9	0.77	0.91	0.86	26Ctl	0.33	0.097	0.3	0.31
05PD	9	2	2.75	0.84	0.58	0.61	0.55	27Ctl	0.35	0.32	0.39	0.27
06PD	23	2	6.8	0.98	0.77	0.83	0.82	28Ctl	0.45	0.26	0.33	0.17
07PD	23	2.5	4.8	0.84	0.77	0.63	0.69	29Ctl	0.14	0.098	0.24	0.042
08PD	13	2	1	1	1	1	1	30Ctl	0.41	0.38	0.42	0.37
09PD	27	2.5	1.4	0.69	0.61	0.58	0.61	31Ctl	0.6	0.41	0.39	0.37
10PD	5	1	0.75	0.88	0.93	0.85	0.91	32Ctl	0.64	0.49	0.52	0.47
11PD	21	2.5	3.25	0.98	0.96	0.92	0.87	33Ctl	0.023	0.042	0	0
12PD	20	2.5	1.8	0.89	0.76	0.75	0.77	34Ctl	0.27	0.3	0.31	0.16
13PD	11	2	3	0.8	0.68	0.75	0.68	35Ctl	0	0	0.22	0.081
14PD	17	2.5	4	0.78	0.65	0.69	0.61	36Ctl	0.53	0.37	0.36	0.23
15PD	29	3	2.9	0.75	0.69	0.68	0.67					
16PD	25	2	0.5	0.59	0.48	0.45	0.46					
17PD	10	1.5	0.5	0.36	0.26	0.24	0.16					
18PD	11	2	0.75	0.34	0.37	0.41	0.35					
19PD	13	2	1.75	0.87	0.81	0.86	0.9					
20PD	19	2.5	3.25	0.47	0.5	0.42	0.46					

BIBLIOGRAPHY

- [1] E. Dorsey and R Constantinescu, "Projected number of people with Parkinson's disease in the most populous nations, 2005 through 2030," *Neurology*, 2007.
- [2] F. Chan, I. T. Armstrong, G. Pari, R. J. Riopelle, and D. P. Munoz, "Deficits in saccadic eye-movement control in parkinson's disease," *Neuropsychologia*, vol. 43, no. 5, pp. 784–796, 2005.
- [3] W. R. Adams, "High-accuracy detection of early parkinson's disease using multiple characteristics of finger movement while typing," *PloS one*, vol. 12, no. 11, e0188226, 2017.
- [4] T.-B. Ahn, B. Na, and D Lee, "Visuospatial dysfunction in cube copying test and clock drawing test in non-demented parkinson's disease: A volumetric analysis," *Parkinsonism & Related Disorders*, vol. 46, e50, 2018.
- [5] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Decision support framework for parkinson's disease based on novel handwriting markers," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 3, pp. 508–516, 2015.
- [6] J Kegl, H Cohen, and H Poizner, "Articulatory consequences of Parkinson's disease: perspectives from two modalities.," *Brain and Cognition*, vol. 40, no. 2, pp. 355–86, 1999.
- [7] J Parkinson, "An essay on the shaking palsy [facsimile: J neuropsychiatry clin neurosc 14: 2 2002]," *London: Sherwood, Neely & Jones*, 1817.
- [8] W. P. Gilks, P. M. Abou-Sleiman, S. Gandhi, S. Jain, A. Singleton, A. J. Lees, K. Shaw, K. P. Bhatia, V. Bonifati, N. P. Quinn, *et al.*, "A common lrrk2 mutation in idiopathic parkinson's disease," *The Lancet*, vol. 365, no. 9457, pp. 415–416, 2005.
- [9] H. Liou, M. Tsai, C. Chen, J. Jeng, Y. Chang, S. Chen, and R. Chen, "Environmental risk factors and parkinson's disease a case-control study in taiwan," *Neurology*, vol. 48, no. 6, pp. 1583–1588, 1997.
- [10] L. Klingelhoefer and H. Reichmann, "Pathogenesis of parkinson disease—the gut–brain axis and environmental factors," *Nature Reviews Neurology*, vol. 11, no. 11, p. 625, 2015.
- [11] M. A. Hernan, B. Takkouche, F. Caamano-Isorna, and J. J. Gestal-Otero, "A meta-analysis of coffee drinking, cigarette smoking, and the risk of parkinson's disease," *Annals of neurology*, vol. 52, no. 3, pp. 276–284, 2002.

- [12] A. Ascherio and M. A. Schwarzschild, "The epidemiology of parkinson's disease: Risk factors and prevention," *The Lancet Neurology*, vol. 15, no. 12, pp. 1257–1272, 2016.
- [13] R. Savica, W. A. Rocca, and J. E. Ahlskog, "When does parkinson disease start?" *Archives of neurology*, vol. 67, no. 7, pp. 798–801, 2010.
- [14] A. Lee and R. M. Gilbert, "Epidemiology of parkinson disease," *Neurologic clinics*, vol. 34, no. 4, pp. 955–965, 2016.
- [15] C. A. Haaxma, B. R. Bloem, G. F. Borm, W. J. Oyen, K. L. Leenders, S. Eshuis, J. Booij, D. E. Dluzen, and M. W. Horstink, "Gender differences in parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 78, no. 8, pp. 819–824, 2007.
- [16] L. S. Ishihara, A. Cheesbrough, C. Brayne, and A. Schrag, "Estimated life expectancy of parkinson's patients compared with the uk population," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 78, no. 12, pp. 1304–1309, 2007.
- [17] R. F. Pfeiffer, Z. K. Wszolek, and M. Ebadi, *Parkinson's Disease*. CRC Press, 2013.
- [18] A. J. Hughes, S. E. Daniel, Y. Ben-Shlomo, and A. J. Lees, "The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service," *Brain*, vol. 125, no. 4, pp. 861–870, 2002.
- [19] D. J. Gelb, E. Oliver, and S. Gilman, "Diagnostic criteria for parkinson disease," *Archives of neurology*, vol. 56, no. 1, pp. 33–39, 1999.
- [20] M. M. Hoehn and M. D. Yahr, "Parkinsonism onset, progression, and mortality," *Neurology*, vol. 17, no. 5, pp. 427–427, 1967.
- [21] A. Hughes, S. Daniel, and A. Lees, "The clinical features of parkinson's disease in 100 histologically proven cases," *Advances in neurology*, vol. 60, p. 595, 1993.
- [22] E. D. Louis, G. Levy, L. J. Cote, H. Mejia, S. Fahn, and K. Marder, "Clinical correlates of action tremor in parkinson disease," *Archives of neurology*, vol. 58, no. 10, pp. 1630–1634, 2001.
- [23] E. D. Louis, L. A. Klatka, Y. Liu, and S. Fahn, "Comparison of extrapyramidal features in 31 pathologically confirmed cases of diffuse lewy body disease and 34 pathologically confirmed cases of parkinson's disease," *Neurology*, vol. 48, no. 2, pp. 376–380, 1997.
- [24] F. B. Horak, D. Dimitrova, and J. G. Nutt, "Direction-specific postural instability in subjects with parkinson's disease," *Experimental neurology*, vol. 193, no. 2, pp. 504–521, 2005.
- [25] S. Fahn, *Recent developments in Parkinson's Disease*. Raven Pr, 1986.

- [26] C. G. Goetz, W. Poewe, O. Rascol, C. Sampaio, G. T. Stebbins, C. Counsell, N. Giladi, R. G. Holloway, C. G. Moore, G. K. Wenning, *et al.*, "Movement disorder society task force report on the hoehn and yahr staging scale: Status and recommendations the movement disorder society task force on rating scales for parkinson's disease," *Movement disorders*, vol. 19, no. 9, pp. 1020–1028, 2004.
- [27] A. Rajput, B Rozdilsky, and A. Rajput, "Accuracy of clinical diagnosis in parkinsonism—a prospective study," *Canadian Journal of Neurological Sciences*, vol. 18, no. 3, pp. 275–278, 1991.
- [28] R. Bhidayasiri and D. Tarsy, "Parkinson's disease: "on-off" phenomenon," in *Movement Disorders: A Video Atlas*, Springer, 2012, pp. 14–15.
- [29] A. Lees, "The on-off phenomenon.," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 52, no. Suppl, pp. 29–37, 1989.
- [30] B. S. Connolly and A. E. Lang, "Pharmacological treatment of parkinson disease: A review," *Jama*, vol. 311, no. 16, pp. 1670–1683, 2014.
- [31] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential Diagnostic Patterns of Dysarthria," *Journal of Speech Language and Hearing Research*, vol. 12, no. 2, p. 246, 1969.
- [32] H Ackermann and W Ziegler, "Articulatory deficits in parkinsonian dysarthria: an acoustic analysis," *J Neurol Neurosurg Psychiatry*, vol. 54, no. 12, pp. 1093–1098, 1991.
- [33] P. Blanchet and G. Snyder, "Speech Rate Deficits in Individuals with Parkinson's Disease: A Review of the Literature," *Journal of Medical Speech - Language Pathology*, vol. 17, no. 1, pp. 1–7, 2009.
- [34] S Skodda, W Grönheit, N Mancinelli, and U Schlegel, "Progression of voice and speech impairment in the course of parkinson's disease: A longitudinal study," *Parkinson's Disease*, vol. 2013, 2013.
- [35] A. B. Walsh, "Basic Parameters of Articulatory Movements and Acoustics in Individuals with Parkinson's Disease," *Movement Disorders*, vol. 27, no. 7, pp. 843–850, 2012.
- [36] K. Forrest, G. Weismer, and G. S. Turner, "Kinematic, acoustic, and perceptual analyses of connected speech produced by parkinsonian and normal geriatric adults," *The Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2608–2622, 1989.
- [37] C. J. Hunker, J. H. Abbs, and S. M. Barlow, "The relationship between parkinsonian rigidity and hypokinesia in the orofacial system a quantitative analysis," *Neurology*, vol. 32, no. 7, pp. 749–749, 1982.

- [38] A. M. Goberman and C. Coelho, "Acoustic analysis of parkinsonian speech i: Speech characteristics and l-dopa therapy," *NeuroRehabilitation*, vol. 17, no. 3, pp. 237–246, 2002.
- [39] J. Rusz, R. Cmejla, H. Ruvzivckova, J. Klempivir, V. Majerova, J. Picmausova, J. Roth, and E. Ruvzivcka, "Acoustic assessment of voice and speech disorders in parkinson's disease through quick vocal test," *Movement Disorders*, vol. 26, no. 10, pp. 1951–1952, 2011.
- [40] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's Disease.," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011, ISSN: 00014966.
- [41] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.
- [42] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of Parkinson's disease from speech," *Computer Speech and Language*, vol. 29, no. 1, pp. 172–185, 2015.
- [43] J. Tetrud, "Preclinical Parkinson's Disease Detection of motor and nonmotor manifestations," *Neurology*, 1991.
- [44] G. Weismer, "Philosophy of research in motor speech disorders," *Clinical Linguistics & Phonetics*, vol. 20, no. 5, pp. 315–349, 2006.
- [45] M. J. Hammer and S. M. Barlow, "Laryngeal somatosensory deficits in parkinson's disease: Implications for speech respiratory and phonatory control," *Experimental brain research*, vol. 201, no. 3, pp. 401–409, 2010.
- [46] S. Scott and F. Caird, "Speech therapy for parkinson's disease.," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 46, no. 2, pp. 140–144, 1983.
- [47] J. Müller, G. K. Wenning, M. Verny, A. McKee, K. R. Chaudhuri, and e. a. Jellinger, "Progression of Dysarthria and Dysphagia in Postmortem-Confirmed Parkinsonian Disorders," *Archives of Neurology*, vol. 58, no. 2, p. 259, 2001.
- [48] 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.
- [49] L. Hartelius and P. Svensson, "Speech and swallowing symptoms associated with parkinson's disease and multiple sclerosis: A survey," *Folia Phoniatrica et Logopaedica*, vol. 46, no. 1, pp. 9–17, 1994.

- [50] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with parkinson's disease," *Behavioural neurology*, vol. 11, no. 3, pp. 131–137, 1999.
- [51] G. Defazio, M. Guerrieri, D. Liuzzi, A. F. Gigante, and V. di Nicola, "Assessment of voice and speech symptoms in early Parkinson's Disease by the Robertson dysarthria profile," *Neurological Sciences*, vol. 37, no. 3, pp. 443–449, 2016.
- [52] S. J. Robertson and F. Thomson, "Speech therapy in parkinson's disease: A study of the efficacy and long term effects of intensive treatment," *International Journal of Language & Communication Disorders*, vol. 19, no. 3, pp. 213–224, 1984.
- [53] S. Sapir, A Pawlas, L Ramig, S Countryman, C O BRIEN, M Hoehn, and L. Thompson, "Speech and voice abnormalities in Parkinson's disease: relation to severity of motor impairment, duration of disease, medication, depression, gender and age," *NCVS Status and Progress Report*, vol. 14, pp. 149–161, 1999.
- [54] C. A. Baumgartner, S. Sapir, and L. O. Ramig, "Voice quality changes following phonatory-respiratory effort treatment (lsvt®) versus respiratory effort treatment for individuals with parkinson disease," *Journal of Voice*, vol. 15, no. 1, pp. 105–114, 2001.
- [55] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of intensive voice treatment (the lee silverman voice treatment [lsvt]) on vowel articulation in dysarthric individuals with idiopathic parkinson disease: Acoustic and perceptual findings," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 4, pp. 899–912, 2007.
- [56] G. S. Berke, B. Gerratt, J. Kreiman, and K. Jackson, "Treatment of parkinson hypophonia with percutaneous collagen augmentation," *The Laryngoscope*, vol. 109, no. 8, pp. 1295–1299, 1999.
- [57] N. K.K., Z. H., and T. H.R., "Speech defects of parkinsonian patients. Effects of levodopa therapy on speech intelligibility," *Neurology*, vol. 23, no. 8, pp. 865–870, 1973.
- [58] J. Rusz, R. Cmejla, H. Ruzivckova, J. Klempivř, V. Majerova, J. Picmausova, J. Roth, and E. Ruvzivcka, "Evaluation of speech impairment in early stages of parkinson's disease: A prospective study with the role of pharmacotherapy," *Journal of Neural Transmission*, vol. 120, no. 2, pp. 319–329, 2013.
- [59] V. Wolfe, J. Garvin, M Bacon, and W Waldrop, "Speech changes in parkinson's disease during treatment with l-dopa," *Journal of communication disorders*, vol. 8, no. 3, pp. 271–279, 1975.

- [60] C. E. Quagliari and G. G. Celesia, "Effect of thalamotomy and levodopa therapy on the speech of parkinson patients," *European neurology*, vol. 15, no. 1, pp. 34–39, 1977.
- [61] P. Svensson, C. Henningson, and S. Karlsson, "Speech motor control in parkinson's disease: A comparison between a clinical assessment protocol and a quantitative analysis of mandibular movements," *Folia Phoniatica et Logopaedica*, vol. 45, no. 4, pp. 157–164, 1993.
- [62] K Larson, L. Ramig, and R. Scherer, "Acoustic and glottographic voice analysis during drug-related fluctuations in parkinson's disease," *Journal of Medical Speech-Language Pathology*, vol. 2, no. 3, pp. 228–239, 1994.
- [63] A. K. Ho, J. L. Bradshaw, and R. Iansek, "For better or worse: The effect of levodopa on speech in parkinson's disease," *Movement Disorders*, vol. 23, no. 4, pp. 574–580, 2008.
- [64] A. Goberman, C. Coelho, and M. Robb, "Phonatory characteristics of parkinsonian speech before and after morning medication: The on and off states," *Journal of communication disorders*, vol. 35, no. 3, pp. 217–239, 2002.
- [65] S. Skodda, W. Visser, and U. Schlegel, "Short-and long-term dopaminergic effects on dysarthria in early parkinson's disease," *Journal of Neural Transmission*, vol. 117, no. 2, pp. 197–205, 2010.
- [66] "Relative fundamental frequency during vocal onset and offset in older speakers with and without Parkinson's disease a)," *Journal of the Acoustic Society of America*, vol. 133, no. 3, pp. 1637–1643, 2012.
- [67] K Chenausky, J MacAuslan, and R Goldhor, "Acoustic analysis of PD speech," *Parkinson's Disease*, 2011.
- [68] M. De Letter, P. Santens, M. De Bodt, G. Van Maele, J. Van Borsel, and P. Boon, "The effect of levodopa on respiration and word intelligibility in people with advanced parkinson's disease," *Clinical neurology and neurosurgery*, vol. 109, no. 6, pp. 495–500, 2007.
- [69] M. De Letter, P. Santens, I. Estercam, G. Van Maele, M. De Bodt, P. Boon, and J. Van Borsel, "Levodopa-induced modifications of prosody and comprehensibility in advanced parkinson's disease as perceived by professional listeners," *Clinical Linguistics & Phonetics*, vol. 21, no. 10, pp. 783–791, 2007.
- [70] C. D. Marsden and J. Parkes, "On-off effects in patients with parkinson's disease on chronic levodopa therapy," *Lancet*, vol. 1, no. 7954, pp. 292–296, 1976.
- [71] E. Critchley, "Peak-dose dysphonia in parkinsonism," *The Lancet*, vol. 307, no. 7958, p. 544, 1976.

- [72] H. A. Kaplan, S. Machover, and A. Rabiner, "A study of the effectiveness of drug therapy in parkinsonism," *The Journal of nervous and mental disease*, vol. 119, no. 5, pp. 398–411, 1954.
- [73] M. Sarno, "Speech impairment in parkinson's disease," *Archives of physical medicine and rehabilitation*, vol. 49, no. 5, pp. 269–275, 1968.
- [74] F. L. Darley, A. E. Aronson, and J. R. Brown, "Clusters of deviant speech dimensions in the dysarthrias," *Journal of speech and hearing research*, vol. 12, no. 3, pp. 462–496, 1969.
- [75] H. Ackermann, I. Hertrich, and T. Hehr, "Oral diadochokinesis in neurological dysarthrias," *Folia Phoniatrica et Logopaedica*, vol. 47, no. 1, pp. 15–23, 1995.
- [76] M. Hirano, "Clinical examination of voice," *Disorders of human communication*, vol. 5, pp. 1–99, 1981.
- [77] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *American Journal of Speech-Language Pathology*, vol. 18, no. 2, pp. 124–132, 2009.
- [78] K. Elemetrics, "Multi-Dimensional Voice Program (MDVP). [Computer program.]," *Pine Brook, NJ: Author*, 1993.
- [79] P. Boersma, "Praat: doing phonetics by computer," <http://www.praat.org/>, 2006.
- [80] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, ACM, 2010, pp. 1459–1462.
- [81] D. G. Hanson, B. R. Gerratt, and P. H. Ward, "Cinegraphic observations of laryngeal function in parkinson's disease," *The Laryngoscope*, vol. 94, no. 3, pp. 348–353, 1984.
- [82] K. S. Perez, L. O. Ramig, M. E. Smith, and C. Dromey, "The parkinson larynx: Tremor and videostroboscopic findings," *Journal of Voice*, vol. 10, no. 4, pp. 354–361, 1996.
- [83] J. Jiang, T. O'Mara, H.-J. Chen, J. I. Stern, D. Vlagos, and D. Hanson, "Aerodynamic measurements of patients with parkinson's disease," *Journal of Voice*, vol. 13, no. 4, pp. 583–591, 1999.
- [84] I. Midi, M. Dogan, M. Koseoglu, G. Can, M. Sehitoglu, and D. Gunal, "Voice abnormalities and their relation with motor dysfunction in parkinson's disease," *Acta Neurologica Scandinavica*, vol. 117, no. 1, pp. 26–34, 2008.
- [85] L. A. Ramig, I. R. Titze, R. C. Scherer, and S. P. Ringel, "Acoustic analysis of voices of patients with neurologic disease: Rationale and preliminary data," *Annals of Otology, Rhinology & Laryngology*, vol. 97, no. 2, pp. 164–172, 1988.

- [86] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, and V. Woisard, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," *European Archives of Oto-rhino-laryngology*, vol. 258, no. 2, pp. 77–82, 2001.
- [87] F. J. Jiménez-Jiménez, J. Gamboa, A. Nieto, J. Guerrero, M. Orti-Pareja, J. A. Molina, E. García-Albea, and I. Cobeta, "Acoustic voice analysis in untreated patients with parkinson's disease," *Parkinsonism & Related Disorders*, vol. 3, no. 2, pp. 111–116, 1997.
- [88] J. H. Rhonda, J. M. Oates, J. P. Debbie, and J. H. Andrew, "Voice characteristics in the progression of Parkinson's disease," *International Journal of Language and Communication Disorders*, vol. 35, no. 3, pp. 407–418, 2000.
- [89] D. A. Rahn, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory impairment in parkinson's disease: Evidence from nonlinear dynamic analysis and perturbation analysis," *Journal of Voice*, vol. 21, no. 1, pp. 64–71, 2007.
- [90] Y. Tanaka, M. Nishio, and S. Niimi, "Vocal acoustic characteristics of patients with parkinson's disease," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 5, pp. 223–230, 2011.
- [91] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "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.
- [92] I. Hertrich and H. Ackermann, "Gender-specific vocal dysfunctions in parkinson's disease: Electrolottographic and acoustic analyses," *Annals of Otology, Rhinology & Laryngology*, vol. 104, no. 3, pp. 197–202, 1995.
- [93] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in parkinson's disease," *Journal of Voice*, vol. 25, no. 4, pp. 467–472, 2011.
- [94] Y. Bang, K Min, Y. Sohn, and S. Cho, "Acoustic characteristics of vowel sounds in patients with Parkinson's Disease," *NeuroRehabilitation*, 2013.
- [95] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "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.

- [96] N. Roy, S. L. Nissen, C. Dromey, and S. Sapir, "Articulatory changes in muscle tension dysphonia: Evidence of vowel space expansion following manual circumlaryngeal therapy," *Journal of communication disorders*, vol. 42, no. 2, pp. 124–135, 2009.
- [97] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, *et al.*, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE transactions on biomedical engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.
- [98] A. Tsanas and M. Little, "Accurate telemonitoring of Parkinson's Disease progression by noninvasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.
- [99] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.
- [100] C. G. Goetz, G. T. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Elble, M. Hallett, J. Nutt, L. Ramig, T. Sanger, *et al.*, "Testing objective measures of motor impairment in early parkinson's disease: Feasibility study of an at-home testing device," *Movement Disorders*, vol. 24, no. 4, pp. 551–556, 2009.
- [101] J. Mekyska, Z. Galaz, Z. Mzourek, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, D. Berankova, M. Faundez-Zanuy, *et al.*, "Assessing progress of parkinson's disease using acoustic analysis of phonation," in *Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference on*, IEEE, 2015, pp. 111–118.
- [102] A. Benba, A. Jilbab, and A. Hammouch, "VOICE ANALYSIS FOR DETECTING PERSONS WITH PARKINSON'S DISEASE USING PLP AND VQ," *Journal of Theoretical & Applied Information Technology*, vol. 70, no. 3, 2014.
- [103] —, "Detecting patients with parkinson's disease using mel frequency cepstral coefficients and support vector machines," *International Journal on Electrical Engineering and Informatics*, vol. 7, no. 2, p. 297, 2015.
- [104] —, "Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients with Parkinson's disease and healthy people," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 449–456, 2016.
- [105] J. A. Logemann and H. B. Fisher, "Vocal Tract Control in Parkinson's Disease," *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, p. 348, 1981.

- [106] G. Weismer and M McNeil, "Articulatory characteristics of parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal-supraglottal coordination," *The dysarthrias: Physiology, acoustics, perception, management*, pp. 101–130, 1984.
- [107] J. A. Robbins, J. A. Logemann, and H. S. Kirshner, "Swallowing and speech production in parkinson's disease," *Annals of neurology*, vol. 19, no. 3, pp. 283–287, 1986.
- [108] E. Q. Wang, L. V. Metman, R. A. Bakay, J. Arzbaecher, B. Bernard, and D. M. Corcos, "Hemisphere-specific effects of subthalamic nucleus deep brain stimulation on speaking rate and articulatory accuracy of syllable repetitions in parkinson's disease," *Journal of medical speech-language pathology*, vol. 14, no. 4, p. 323, 2006.
- [109] J. Godino-Llorente, S Shattuck-Hufnagel, J. Choi, L Moro-Velázquez, and J. Gómez-García, "Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers," *PloS one*, vol. 12, no. 12, e0189583, 2017.
- [110] B. T. Harel, M. S. Cannizzaro, H. Cohen, N. Reilly, and P. J. Snyder, "Acoustic characteristics of Parkinsonian speech: A potential biomarker of early disease progression and treatment," *Journal of Neurolinguistics*, vol. 17, no. 6, pp. 439–453, 2004.
- [111] S. Gräber, I. Hertrich, I. Daum, S. Spieker, and H. Ackermann, "Speech perception deficits in Parkinson's disease: Underestimation of time intervals compromises identification of durational phonetic contrasts," *Brain and Language*, vol. 82, no. 1, pp. 65–74, 2002.
- [112] K. Tjaden, J. Lam, and G. Wilding, "Vowel acoustics in Parkinson's Disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions.," *Journal of speech, language, and hearing research : JSLHR*, vol. 56, no. 5, pp. 1485–502, 2013.
- [113] K. Tjaden, V. Martel-Sauvageau, D. J. R., B. D. R., G. L. A., M. R. M., and S. M., "Consonant Acoustics in Parkinson's Disease and Multiple Sclerosis: Comparison of Clear and Loud Speaking Conditions," *American Journal of Speech-Language Pathology*, vol. 26, no. 2S, p. 569, 2017.
- [114] S. Skodda, W. Grönheit, and U. Schlegel, "Impairment of vowel articulation as a possible marker of disease progression in parkinson's disease," *PloS one*, vol. 7, no. 2, e32132, 2012.
- [115] J. Whitfield and A. Goberman, "Articulatory acoustic vowel space: Application to clear speech in individuals with Parkinson's disease," *Journal of Communication Disorders*, 2014.

- [116] K. Forrest, G. Weismer, P. Milenkovic, and R. N. Dougall, "Statistical analysis of word-initial voiceless obstruents: Preliminary data," *The Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 115–123, 1988.
- [117] K. J. Reilly and K. A. Spencer, "Sequence complexity effects on speech production in healthy speakers and speakers with hypokinetic or ataxic dysarthria," *PloS one*, vol. 8, no. 10, 2013.
- [118] M. P. Caligiuri, "The influence of speaking rate on articulatory hypokinesia in parkinsonian dysarthria," *Brain and Language*, vol. 36, no. 3, pp. 493–502, 1989.
- [119] K. Forrest and G. Weismer, "Dynamic aspects of lower lip movement in parkinsonian and neurologically normal geriatric speakers' production of stress," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 2, pp. 260–272, 1995.
- [120] Y. Kim, G. Weismer, R. D. Kent, and J. R. Duffy, "Statistical models of F2 slope in relation to severity of dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 6, pp. 329–335, 2009.
- [121] S Sapir, L. Ramig, J. Spielman, and C Fox, "Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech," *Journal of Speech, Language, and ...*, 2010.
- [122] M. Novotny, J. Rusz, R. mejla, and E. Rka, "Automatic evaluation of articulatory disorders in parkinson s disease," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1366–1378, 2014.
- [123] J. R. Orozco-Arroyave, F Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, K Daqrouq, S Skodda, J Rusz, and E Nöth, "Automatic detection of Parkinson's disease in running speech spoken in three different languages.," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, p. 481, 2016.
- [124] J. R. Orozco-Arroyave, J. Vasquez-Correa, F. Honig, J. D. Arias-Londono, J. Vargas-Bonilla, S. Skodda, J. Rusz, and E Noth, "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP, IEEE*, 2016, pp. 6490–6494.
- [125] T. Khan, J. Westin, and M. Dougherty, "Cepstral separation difference: A novel approach for speech impairment quantification in Parkinson's disease," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 25–34, 2014.
- [126] S. U. Hahm and J. U. Wang, "Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, International Speech and Communication Association*, vol. 2015, 2015.

- [127] J. Orozco-Arroyave and J. Arias-Londoño, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease.," *Proceedings on the International Conference on Language Resources and Evaluation (LREC)*, 2014.
- [128] T. Grosz, R. Busa-Fekete, and G. Gosztolya, "Assessing the Degree of Nativeness and Parkinson's Condition Using Gaussian Processes and Deep Rectifier Neural Networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 919–923.
- [129] G. An, D. G. Brizan, M. Ma, M. Morales, A. R. Syed, and A. Rosenberg, "Automatic Recognition of Unified Parkinson's Disease Rating from Speech with Acoustic, i-Vector and Phonotactic Features," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 508–512.
- [130] S. David, G. Kiss, and K. Vicsi, "Estimating the Severity of Parkinson's Disease from Speech Using Linear Regression and Database Partitioning," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.
- [131] A. Zlotnik, J. M. Montero, R. San-Segundo, and A. Gallardo-Antolin, "Random forest-based prediction of Parkinson's disease progression using acoustic, ASR and intelligibility features," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.
- [132] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vásquez-Correa, and E. Nöth, "Characterisation of voice quality of Parkinson's disease using differential phonological posterior features," *Computer Speech and Language*, vol. 46, pp. 196–208, 2017.
- [133] G. Monrad-Krohn, "The third element of speech: Prosody and its disorders," *Problems in dynamic neurology*, pp. 101–118, 1963.
- [134] G. J. Canter, "Speech characteristics of patients with Parkinson's disease: I. intensity, pitch and duration," *The Journal of Speech and Hearing Disorders*, vol. 28, no. 3, pp. 221–229, 1963.
- [135] J. Illes, E. Metter, W. Hanson, and S. Iritani, "Language production in Parkinson's Disease: Acoustic and linguistic considerations," *Brain and Language*, vol. 33, no. 1, pp. 146–160, 1988.
- [136] 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.

- [137] S. Skodda, W. Grö, U. Schlegel, W. Grönheit, and U. Schlegel, "Intonation and Speech Rate in Parkinson's Disease: General and Dynamic Aspects and Responsiveness to Levodopa Admission," *Journal of Voice*, vol. 25, no. 4, e199–205, 2011.
- [138] G. Le Dorze, J. Ryalls, C. Brassard, N. Boulanger, and D. Ratte, "A comparison of the prosodic characteristics of the speech of people with parkinson's disease and friedreich's ataxia with neurologically normal speakers," *Folia Phoniatrica et Logopaedica*, vol. 50, no. 1, pp. 1–9, 1998.
- [139] J. Mobes, G. Joppich, F. Stiebritz, R. Dengler, and C. Schroder, "Emotional speech in Parkinson's disease," *Movement Disorders*, vol. 23, no. 6, pp. 824–829, 2008.
- [140] B. T. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's Disease: A longitudinal case study," *Brain and Cognition*, vol. 56, no. 1, pp. 24–29, 2004.
- [141] B. Teston, "L'évaluation objective des dysfonctionnements de la voix et de la parole; 2e partie: les dysphonies," *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, vol. 20, pp. 169–232, 2001.
- [142] C. E. Stepp, "Relative fundamental frequency during vocal onset and offset in older speakers with and without Parkinson's disease," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1637–1643, 2013.
- [143] A. Ho, J. Bradshaw, R. Iansek, and R. Alfredson, "Speech volume regulation in Parkinson's disease: Effects of implicit cues and explicit instructions," *Neuropsychologia*, 1999.
- [144] E. Lombard, "Le signe de l'élevation de la voix," *Ann Maladies de L'Oreille et du Larynx*, vol. 37, p. 2, 1911.
- [145] C. M. Fox and L. O. Ramig, "Vocal Sound Pressure Level and Self-Perception of Speech and Voice in Men and Women with Idiopathic Parkinson Disease," *American Journal of Speech-Language Pathology*, vol. 6, no. 2, pp. 85–94, 1997.
- [146] E. J. Metter and W. R. Hanson, "Clinical and acoustical variability in hypokinetic dysarthria," *Journal of communication disorders*, vol. 19, no. 5, pp. 347–366, 1986.
- [147] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Acoustic analysis in the differentiation of parkinson's disease and major depression," *Journal of Psycholinguistic Research*, vol. 21, no. 5, pp. 383–399, 1992.

- [148] A. Bandini, F. Giovannelli, S. Orlandi, S. Barbagallo, M. Cincotta, P. Vanni, R. Chiaramonti, A. Borgheresi, G. Zaccara, and C. Manfredi, "Automatic identification of dysprosody in idiopathic parkinson's disease," *Biomedical Signal Processing and Control*, vol. 17, pp. 47–54, 2015.
- [149] C. L. Ludlow, N. P. Connor, and C. J. Bassich, "Speech timing in parkinson's and huntington's disease," *Brain and language*, vol. 32, no. 2, pp. 195–214, 1987.
- [150] S. Skodda and U. Schlegel, "Speech rate and rhythm in parkinson's disease," *Movement Disorders*, vol. 23, no. 7, pp. 985–992, 2008.
- [151] J. Rusz, J. Hlavnicka, R. Cmejla, and E. Ruzicka, "Automatic evaluation of speech rhythm instability and acceleration in dysarthrias associated with basal ganglia dysfunction," *Frontiers in bioengineering and biotechnology*, vol. 3, 2015.
- [152] H. Ackermann, J. Konczak, and I. Hertrich, "The temporal control of repetitive articulatory movements in parkinson's disease," *Brain and language*, vol. 56, no. 2, pp. 312–319, 1997.
- [153] S. Skodda, "Aspects of speech rate and regularity in Parkinson's disease," *Journal of the Neurological Sciences*, vol. 310, no. 1–2, pp. 231–236, 2011.
- [154] K. Bunton, "Patterns of lung volume use during an extemporaneous speech task in persons with Parkinson disease," *Journal of Communication Disorders*, vol. 38, no. 5, pp. 331–348, 2005.
- [155] A. W. Darkins, V. A. Fromkin, and D. F. Benson, "A characterization of the prosodic loss in Parkinson's disease," *Brain and Language*, vol. 34, no. 2, pp. 315–327, 1988.
- [156] M. D. Pell, "On the Receptive Prosodic Loss in Parkinson's Disease," *Cortex*, vol. 32, no. 4, pp. 693–704, 1996.
- [157] A. J. Lloyd, "Comprehension of Prosody in Parkinson's Disease," *Cortex*, vol. 35, no. 3, pp. 389–402, 1999.
- [158] M. D. Pell and C. L. Leonard, "Processing emotional tone from speech in Parkinson's disease: a role for the basal ganglia," *Cogn Affect Behav Neurosci*, vol. 3, no. 4, pp. 275–288, 2003.
- [159] C. Schroder, J. Mobes, M. Schutze, F. Szymanowski, W. Nager, M. Bangert, T. F. Münte, and R. Dengler, "Perception of emotional speech in parkinson's disease," *Movement Disorders*, vol. 21, no. 10, pp. 1774–1778, 2006.
- [160] 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.

- [161] N. Miller, E. Noble, D. Jones, L. Allcock, and D. J. Burn, "How do i sound to me? perceived changes in communication in parkinson's disease," *Clinical rehabilitation*, vol. 22, no. 1, pp. 14–22, 2008.
- [162] Y. Bocanegra, A. M. García, D. Pineda, O. Buriticá, A. Villegas, F. Lopera, D. Gómez, C. Gómez-Arias, J. F. Cardona, N. Trujillo, *et al.*, "Syntax, action verbs, action semantics, and object semantics in parkinson's disease: Dissociability, progression, and executive influences," *Cortex*, vol. 69, pp. 237–254, 2015.
- [163] J. Illes, "Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, huntington's, and parkinson's," *Brain and language*, vol. 37, no. 4, pp. 628–642, 1989.
- [164] T. Benke, C Hohenstein, W Poewe, and B Butterworth, "Repetitive speech phenomena in parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 69, no. 3, pp. 319–324, 2000.
- [165] A. M. Goberman, M. Blomgren, and E. Metzger, "Characteristics of speech disfluency in parkinson disease," *Journal of Neurolinguistics*, vol. 23, no. 5, pp. 470–478, 2010.
- [166] D. Van Lancker Sidtis, K. Cameron, and J. J. Sidtis, "Dramatic effects of speech task on motor and linguistic planning in severely dysfluent parkinsonian speech.," EN, *Clinical Linguistics & Phonetics*, vol. 26, no. 8, pp. 695–711, 2012.
- [167] B. Walsh and A. Smith, "Linguistic complexity, speech production, and comprehension in parkinson's disease: Behavioral and physiological indices," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 3, pp. 787–802, 2011.
- [168] D Natsopoulos, G Grouios, S Bostantzopoulou, G Mentenopoulos, Z Katsarou, and J Logothetis, "Algorithmic and heuristic strategies in comprehension of complement clauses by patients with parkinson's disease," *Neuropsychologia*, vol. 31, no. 9, pp. 951–964, 1993.
- [169] J. Rodriguez-Ferreiro, M. Menendez, R. Ribacoba, and F. Cuetos, "Action naming is impaired in parkinson disease patients," *Neuropsychologia*, vol. 47, no. 14, pp. 3271–3274, 2009.
- [170] E. Herrera and F. Cuetos, "Action naming in parkinson's disease patients on/off dopamine," *Neuroscience letters*, vol. 513, no. 2, pp. 219–222, 2012.
- [171] R. Matison, R. Mayeux, J. Rosen, and S. Fahn, "'tip-of-the-tongue' phenomenon in parkinson disease," *Neurology*, vol. 32, no. 5, pp. 567–567, 1982.

- [172] D. V. L. Sidtis, J. Choi, A. Alken, and J. J. Sidtis, "Formulaic language in parkinson's disease and alzheimer's disease: Complementary effects of subcortical and cortical dysfunction," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 5, pp. 1493–1507, 2015.
- [173] A. C. Roberts, "Characterizing spoken discourse in individuals with parkinson disease without dementia," 2014.
- [174] M. I. Schlesinger and V. Hlavác, *Ten lectures on statistical and structural pattern recognition*. Springer Science & Business Media, 2013, vol. 24.
- [175] T Bocklet, S Steidl, E Nöth, and S Skodda, "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.
- [176] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.
- [177] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect parkinson's disease," *Applied Soft Computing*, vol. 62, pp. 649–666, 2018.
- [178] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, *et al.*, "Neurospeech: An open-source software for parkinson's speech analysis," *Digital Signal Processing*, 2017.
- [179] P Enderby, "Frenchay dysarthria assessment," *British Journal of Disorders of Communication*, vol. 15, no. 3, pp. 165–173, 1980.
- [180] Z Galaz, Z Mzourek, J Mekyska, and Z Smekal, "Degree of Parkinson's Disease Severity Estimation Based on Speech Signal Processing," *IEEE 39th International*, pp. 503–506, 2016.
- [181] T. Tykalova, J. Ruzs, J. Klempir, R. Cmejla, and E. Ruzicka, "Distinct patterns of imprecise consonant articulation among Parkinson's disease, progressive supranuclear palsy and multiple system atrophy," *Brain and language*, vol. 165, pp. 1–9, 2017.
- [182] L. Ramig, C. Bonitati, J. Lemke, Y Horii, *et al.*, "Voice treatment for patients with parkinson disease: Development of an approach and preliminary efficacy data," *Journal of Medical Speech-Language Pathology*, vol. 2, no. 3, pp. 191–209, 1994.
- [183] T Villa-Canas, J. D. Arias-Londono, J. R. Orozco-Arroyave, J. Vargas-Bonilla, and E. Noth, "Low-frequency components analysis in running speech for the automatic detection of parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

- [184] Y. Kim and Y. Choi, "A Cross-Language Study of Acoustic Predictors of Speech Intelligibility in Individuals With Parkinson's Disease," *Journal of Speech Language and Hearing Research*, vol. 60, no. 9, p. 2506, 2017.
- [185] R. L. Ringel and W. J. Chodzko-Zajko, "Vocal indices of biological age," *Journal of Voice*, vol. 1, no. 1, pp. 31–37, 1987.
- [186] S. E. Linville, "The sound of senescence," *Journal of voice*, vol. 10, no. 2, pp. 190–200, 1996.
- [187] P. Zwirner and G. J. Barnes, "Vocal tract steadiness: A measure of phonatory and upper airway motor control during phonation in dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 35, no. 4, pp. 761–768, 1992.
- [188] J. C. Jimenez-Monsalve, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, and P. Gomez-Vilda, "Phonation and articulation analyses in laryngeal pathologies, cleft lip and palate, and parkinson's disease," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10338 LNCS, 2017, pp. 424–434.
- [189] R. D. Kent and J. C. Rosenbek, "Prosodic disturbance and neurologic lesion," *Brain and Language*, vol. 15, no. 2, pp. 259–291, 1982.
- [190] G. Weismer and J. Wildermuth, "Formant trajectory characteristics in persons with Parkinson, cerebellar, and upper motor neuron disease," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, p. 2892, 1998.
- [191] Y. E. Huh, J. Park, M. K. Suh, S. E. Lee, J. Kim, Y. Jeong, H.-T. Kim, and J. W. Cho, "Differences in early speech patterns between Parkinson variant of multiple system atrophy and Parkinson's disease," *Brain and language*, vol. 147, pp. 14–20, 2015.
- [192] G. Weismer, Y. Kim, B. Maassen, and P. van Lieshout, "Classification and taxonomy of motor speech disorders: What are the issues," *Speech motor control: New developments in basic and applied research*, pp. 229–241, 2010.
- [193] S. Pinto, R. Cardoso, J. Sadat, I. Guimaraes, C. Mercier, H. Santos, C. Atkinson-Clement, J. Carvalho, P. Welby, P. Oliveira, *et al.*, "Dysarthria in individuals with parkinson's disease: A protocol for a binational, cross-sectional, case-controlled study in french and european portuguese (fralusopark)," *BMJ open*, vol. 6, no. 11, e012885, 2016.
- [194] A. J. Hughes, S. E. Daniel, L. Kilford, and A. J. Lees, "Accuracy of clinical diagnosis of idiopathic parkinson's disease: A clinico-pathological study of 100 cases," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 55, no. 3, pp. 181–184, 1992.

- [195] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.
- [196] P. J. Moreno, C. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Fifth International Conference on Spoken Language Processing*, 1998.
- [197] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.
- [198] J. I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, S. Aguilera-Navarro, and P. Gomez-Vilda, "An integrated tool for the diagnosis of voice disorders," *Medical Engineering and Physics*, vol. 28, no. 3, pp. 276–289, 2006.
- [199] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, and C. Nadeu, "Albayzín speech database: Design of the phonetic corpus," in *Eurospeech 1993. Proceedings of the 3rd European Conference on Speech Communication and Technology*, ISCA, vol. 1, 1993, pp. 175–178, ISBN: 1018-4074.
- [200] M. Korvas, O. Platek, O. Dusek, L. Zilka, and F. Jurcicek, "Free english and czech telephone speech corpus shared under the cc-by-sa 3.0 license," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.
- [201] L. Flórez, "El español hablado en colombia y su atlas lingüístico," *Thesaurus*, vol. 1, no. 2, pp. 268–356, 1963.
- [202] J. Kim, M. Nasir, R. Gupta, M. Segbroeck, D. Bone, M. Black, Z. I. Skordilis, Z. Yang, P. Georgiou, and S. Narayanan, "Automatic estimation of parkinson's disease severity from diverse speech tasks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 914–918, 2015.
- [203] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.
- [204] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980, ISSN: 00963518.
- [205] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 10, pp. 1943–1953, 2006.

- [206] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, "Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia)," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 2005)*, ISCA, 2005, pp. 149–152.
- [207] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 36–47, 2008.
- [208] A. Belhaj, A. Bouzid, and N. Ellouze, "Edema and nodule pathological voice identification by svm classifier on speech signal," *International Review on Computers and Software (IRECOS)*, vol. 10, no. 5, pp. 495–501, 2015.
- [209] A. Mesaros and J. Astola, "The mel-frequency cepstral coefficients in the context of singer identification.," in *ISMIR*, Cite-seer, 2005, pp. 610–613.
- [210] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [211] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.
- [212] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.
- [213] "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.
- [214] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.
- [215] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [216] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization.," *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2012.
- [217] M. Li and S. Narayanan, "Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification," *Computer Speech & Language*, vol. 28, no. 4, pp. 940–958, 2014.

- [218] H. Behravan, V. Hautamäki, and T. Kinnunen, "Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish," *Speech Communication*, vol. 66, pp. 118–129, 2015.
- [219] 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 (methodological)*, pp. 1–38, 1977.
- [220] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 249–252.
- [221] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.
- [222] A. Quilis, *Tratado de fonología y fonética españolas*. Editorial Gredos, 1993.
- [223] N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gomez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006.
- [224] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [225] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, "Acoustic studies of dysarthric speech: Methods, progress, and potential," *Journal of communication disorders*, vol. 32, no. 3, pp. 141–186, 1999.
- [226] S Arora, V Venkataraman, A Zhan, S Donohue, K. Biglan, E. Dorsey, and M. Little, "Detecting and monitoring the symptoms of parkinson's disease using smartphones: A pilot study," *Parkinsonism & related disorders*, vol. 21, no. 6, pp. 650–653, 2015.
- [227] M. Shahbakhi, D. T. Far, and E. Tahami, "Speech analysis for diagnosis of parkinson's disease using genetic algorithm and support vector machine," *Journal of Biomedical Science and Engineering*, vol. 7, pp. 147–156, 2014.
- [228] J. Hlavnivcka, R. Cmejla, T. Tykalova, K. Sonka, E. Ruuzicka, and J. Ruzs, "Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Scientific reports*, vol. 7, no. 1, p. 12, 2017.
- [229] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

- [230] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldil," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.
- [231] T. Pruthi, "Analysis, vocal-tract modeling and automatic detection of vowel nasalization," PhD thesis, University of Maryland, College Park, 2007.
- [232] H.-D. Huici, H. A. Kairuz, H. Martens, G. Van Nuffelen, and M. De Bodt, "Speech rate estimation in disordered speech based on spectral landmark detection," *Biomedical Signal Processing and Control*, vol. 27, pp. 1–6, 2016.
- [233] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [234] C. Park, N. F. Chen, Y. Jung, and J. Slifka, "Automatically determining acoustic landmark sequences using physiological constraints," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2008.
- [235] R. J. Baken and R. F. Orlikoff, *Clinical measurement of speech and voice*. Cengage Learning, 2000.
- [236] S. Feijoo and C. Hernández, "Short-term stability measures for the evaluation of vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 324–334, 1990.
- [237] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1329–1334, 1986.
- [238] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.
- [239] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio—a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.
- [240] L. Atlas, P. Clark, and S. Schimmel, *Modulation Toolbox Version 2.1 for MATLAB*, 2010. [Online]. Available: <http://isdl.ee.washington.edu/projects/modulationtoolbox/>.
- [241] S. M. Schimmel and L. E. Atlas, "Coherent envelope detection for modulation filtering of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2005, pp. 221–224.

- [242] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the grbas scale," *BioMed research international*, vol. 2015, 2015.
- [243] L. Moro-Velazquez, J. A. Gomez-Garcia, and J. I. Godino-Llorente, "Voice pathology detection using modulation spectrum-optimized metrics," *Frontiers in bioengineering and biotechnology*, vol. 4, p. 1, 2016.
- [244] R. Elble, "Central mechanisms of tremor," *Journal of clinical neurophysiology*, 1996.
- [245] M. Bové, N. Daamen, C. Rosen, C.-C. Wang, L. Sulica, and J. Gartner-Schmidt, "Development and validation of the vocal tremor scoring system," *The Laryngoscope*, vol. 116, no. 9, pp. 1662–1667, 2006.
- [246] R. Peters and R. Strickland, "Image complexity metrics for automatic target recognizers," *Automatic Target Recognizer System and Technology Conference*, 1990.
- [247] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. Cambridge University Press, 2004.
- [248] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.
- [249] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in non-stationary heartbeat time series," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 5, no. 1, pp. 82–87, 1995.
- [250] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proceedings of the National Academy of Sciences*, vol. 88, no. 6, pp. 2297–2301, 1991.
- [251] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, H2039–H2049, 2000.
- [252] H.-B. Xie, W.-X. He, and H. Liu, "Measuring time series regularity using nonlinear similarity-based sample entropy," *Physics Letters A*, vol. 372, no. 48, pp. 7140–7146, 2008.
- [253] L.-S. Xu, K.-Q. Wang, and L. Wang, "Gaussian kernel approximate entropy algorithm for analyzing irregularity of time-series," in *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, IEEE, vol. 9, 2005, pp. 5605–5608.

- [254] W. Chen, Z. Wang, H. Xie, and W. Yu, "Characterization of surface emg signal based on fuzzy entropy," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 15, no. 2, pp. 266–272, 2007.
- [255] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation entropy and its main biomedical and econophysics applications: A review," *Entropy*, vol. 14, no. 8, pp. 1553–1577, 2012.
- [256] J. D. Arias-Londono, J. I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, and G. Castellanos-Dominguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.
- [257] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1415–1438, 2003.
- [258] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1912–1924, 1999.
- [259] K. Vijayan, V. Kumar, and K. S. R. Murty, "Feature extraction from analytic phase of speech signals for speaker verification," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [260] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP, IEEE*, 2014, pp. 1695–1699.

DECLARATION

I hereby declare that this thesis is my own original work and that, to the best of my knowledge, it contains no materials previously published or written by another person, except where due acknowledgment is made in the thesis. All the assistance received in preparing this thesis has been acknowledged and information derived from the published and unpublished work of others has been acknowledged in the text and referenced.

Madrid,

Laureano Moro Velázquez

ACKNOWLEDGMENT

This work was supported by the Ministry of Economy and Competitiveness of Spain under projects TEC2012-38630-C04-01 (grant BES-2013-062984) and DPI2017-83405-R1.