CONTRIBUTIONS TO THE DESIGN OF AUTOMATIC
VOICE QUALITY ANALYSIS SYSTEMS USING SPEECH
TECHNOLOGIES

JORGE ANDRÉS GÓMEZ GARCÍA, MEng.

Advisor:

JUAN IGNACIO GODINO LLORENTE, PhD.

Thesis submitted to opt for the degree of *Doctor of Philosophy* in:
Communications Systems and Technologies
Tecnologías y Sistemas de Comunicación

Departamento de Señales, Sistemas y Radiocomunicaciones

Escuela Técnica Superior de Ingenieros de Telecomunicación

Universidad Politécnica de Madrid

7th February, 2018

Jorge Andrés Gómez García, MEng.

*Contributions to the Design of Automatic Voice Quality Analysis Systems using
Speech Technologies*

PRESENTED TO OPT FOR THE PHD. DEGREE IN:

Communications Systems and Technologies

Tecnologías y Sistemas de Comunicación

INSTITUTION:

Departamento de Señales, Sistemas y Radiocomunicaciones

Escuela Técnica Superior de Ingenieros de Telecomunicación

Universidad Politécnica de Madrid

ADVISOR:

Juan Ignacio Godino Llorente, PhD.

Madrid

7th February, 2018

TITLE OF THE THESIS:

Contributions to the Design of Automatic Voice Quality Analysis Systems using Speech Technologies

AUTHOR:

Jorge Andrés Gómez García, MEng.

ADVISOR:

Juan Ignacio Godino Llorente, PhD.

COMMITTEE MEMBERS:

Athanasios Tsanas, PhD.	<i>University of Edinburgh</i>
Daniel Ramos Castro, PhD.	<i>Universidad Autónoma de Madrid</i>
José Luis Blanco Murillo, PhD.	<i>Universidad Politécnica de Madrid</i>
Alfonso Ortega Giménez, PhD.	<i>Universidad de Zaragoza</i>
Manuel Blanco Velasco, PhD.	<i>Universidad de Alcalá</i>

EXTERNAL REVIEWERS:

Julián David Arias Londoño, PhD.	<i>Universidad de Antioquia</i>
Andrés Marino Álvarez Meza, PhD.	<i>Universidad Tecnológica de Pereira</i>
Gabriel Alejandro Alzamendi, PhD.	<i>Universidad Nacional de Entre Ríos</i>

Madrid, on the 7th of February, 2018. After the defence of the PhD dissertation at E.T.S.I.S de Telecomunicación of the Universidad Politécnica de Madrid, the board agrees to grant the following qualification:

The chair

The secretary

The members

Dedicada a mis padres Adiela y Rubiél.

In memoriam de Rosalba y Gerardo.

“Deep in the human unconscious is a pervasive need for a logical universe that makes sense. But the real universe is always one step beyond logic.”

Frank Herbert, *Dune*.

AGRADECIMIENTOS

En primer lugar agradezco de manera muy especial a mi director Juan Ignacio Godino Llorente por haberme acompañado durante esta travesía llamada doctorado. Gracias miles por los buenos momentos compartidos durante todos estos años, por el apoyo y consejo constante, por la oportunidad de crecer como investigador y como persona en el seno del laboratorio; pero sobre todo gracias por ser esa persona cercana, afable, paciente y generosa, de quien he podido aprender muchísimo a nivel humano y académico. Sin duda alguna, él es un gran valedor de que esta tesis se haya podido desarrollar. Un agradecimiento gigantesco para Laureano Moro, con quien he tenido la oportunidad de compartir muchos gratos momentos y de crecer durante esta aventura doctoral. Gracias por haber estado siempre en aquellas ocasiones en que nada funcionaba (cosa que era muy frecuente), gracias por haberme ayudado a no perder la cabeza cuando la rutina era omnipresente, y gracias simplemente por estar allí, teniendo tertulias vigoréticas mientras disfrutábamos de unos alfajores junto a Paquito. Gracias enormes Laure.

Gracias muy especiales a Germán Castellanos, uno de los grandes responsable de que hubiese elegido seguir el camino de la ciencia. Gracias por la paciencia, el apoyo, las discusiones, los consejos, y por haber estado siempre cerca, siendo parte fundamental de mi formación. Extiendo también un agradecimiento gigante a Julián David Arias, soporte y guía desde la época de mi máster, y uno de los principales responsables de que mi estancia en España fuese posible. Un agradecimiento enorme para Janaína Mendes por todas las discusiones que hemos tenido y por la paciencia infinita haciendo las evaluaciones perceptuales de las bases de datos, aún cuando esto significase una carrera -casi constante- a contrarreloj. Gracias también a Gustavo, Nico y Víctor, antiguos miembros del laboratorio de Bioingeniería y Optoelectrónica, y con quienes he compartido muchos buenos momentos durante todo este proceso. Gracias a José Luis Blanco, gran investigador y gran persona, y con quien he tenido el grato privilegio de trabajar. Extiendo también las gracias a Gabriel Alzamendi por todo el apoyo recibido y por sus asesorías en temas alfajorísticos.

Gracias a todos esos amigos que en la lejanía y en la cercanía han estado siempre presentes. Doy gracias muy especiales a José Vila, gran amigo y testigo a primera mano de las altas y bajas de todo este proceso. Gracias por la compañía durante los buenos momentos y el apoyo constante durante los malos. Agradezco también a Sebastián García, Juan Pablo Aristizábal y Sady Rojas, grandes e incondicionales amigos, a quienes a pesar de la distancia he sentido siempre cerca. Gracias enormes por el cariño y apoyo que siempre he recibido por su parte. Gracias también a aquellos que han sido parte de todo este proceso: a Jian, a Huang, a Diego Peluffo, a Rodri, a Blanca, a Dani, a Eliana, a Paola, a Estefa y a Claudia.

Durante mi estancia en Canadá tuve la oportunidad de conocer a gente maravillosa, que hicieron de mi estadía una experiencia de vida llenadora. Gracias primeramente a Frank Rudzicz quien me abrió con gran disposición la puerta de su laboratorio, permitiéndome trabajar en temas de investigación nuevos y emocionantes. Gracias a Sin (first name) Tung (last name) por la pasta a la puttanesca, los cafés en Tim Hortons y las manzanas acarameladas. Gracias por las correcciones a la tesis y por ser una gran compañera y amiga durante mi corta pero vibrante estancia canadiense. Gracias a mi parcerera de mil aventuras, Susana, por todos los gratos momentos vividos, y por ayudarme a hacer más llevadero el laundry y ese recalcitrante frío invernal. Gracias también a Irene, a Bryan y a Song por los cafés, las cervezas, los meetups y las mañanas de hiking.

No podría dejar de agradecer a toda mi familia, quienes han sentido en carne propia este doctorado tanto como yo. Gracias infinitas a mi familia española: a Luz, César, Álvaro y César Jr., por haberme abierto la puerta de su casa y de su corazón durante esta experiencia vital. Su soporte y cariño han sido fundamentales para haber podido recorrer este camino. Gracias también a mi familia colombiana: a mi madre Adiela, a mi hermana Viviana, a mi padre Rubiél, a mi sobrina Salomé; a mis tíos Alba Lucía, Diego y Fernando; a mis primos Juan Camilo, Isabella, Juan David, Sebastián y a Johanna. A todos ellos gracias por la paciencia, por el amor recibido, y por hacerse sentirse cerca a pesar de los muchos kilómetros que nos separan.

Gracias también a las personas del grupo de Procesado y Reconocimiento de Señal, a mis compañeros de Aikido, y a todos a quienes han contribuido directa o indirectamente a hacer que esta tesis fuese una realidad.

ABSTRACT

The production of speech relies in a complex process to generate audible outputs for, most typically, communication purposes. Not only speech contains a message encoded in the form of language, but also delivers information about sex, age, condition, and diverse aspects describing the speaker itself. Due to this fact, there exists a great interest in designing systems that extract this non-linguistic information for automatic analysis purposes.

One interesting application -on which this thesis is centred- is in the design of automatic systems capable of characterising the presence and severity of voice disorders. This has potential applications as objective supplementary tools in clinical settings. Notwithstanding, the design of automatic systems poses several problems that include the intrinsic variability of speech, the simultaneous presence of multiple phenomena characterising vocal pathology, the existence of spurious extralinguistic information, or the reliance on perceptual assessments which are highly subjective.

With these antecedents in mind, this thesis evaluates the influence of extralinguistic information, differing types of speech tasks, diverse decision machines and characteristics, in the design of *automatic voice quality analysis* systems whose objective is to generalise decisions about the presence and severity of pathologies present in voices and/or speech. A novel methodology based on feature ranking algorithms, ordinal classification and Gaussian regression is also proposed to emulate the perceptual capabilities of a human evaluator. The regressor is used to convert the discrete perceptual scale to a continuum, more in accordance to the nature of the evaluations. Moreover, the robustness of the proposed systems is evaluated in several cross-database experiments.

Results indicate that the sex of the speaker plays an important role in automatic voice quality analysis systems and that hierarchical designs should be considered. It has also been found that the most consistent set of features for both pathology detection and assessment tasks, are two perturbation measures and a descriptor of the dispersion in *modulation spectra* representations: *glottal-to-noise excitation ratio*, *cepstral harmonics-to-noise ratio* and *rate of points above linear average*. The best automatic detector trained with the *Saarbrücken voice disorders database* achieves an **AUC** of 0.88 when the information provided by the different speech tasks is fused via logistic regression. In several cross-database scenarios, **AUC** varies between 0.75 to 0.94, thus demonstrating the robustness of the system. These are some of the best efficiencies reported in literature using this database. The best assessment system incurs in errors that differ on average half an unit from the actual label, when $\mathcal{G}$ and $\mathcal{B}$ are considered in cross-database settings. Moreover, the system has been assessed clinically by an expert who certified its validity. Results for the system clinically evaluated are of about 0.3 units for the $\mathcal{G}$ trait.

RESUMEN

La producción del habla es un proceso complejo que busca producir señales audibles que son empleadas, generalmente, con fines comunicativos. No solo el habla contiene un mensaje codificado, sino que también entrega información acerca del sexo, la edad, la condición y aspectos que describen al hablante. Debido a esto, existe un gran interés en diseñar sistemas que extraigan esta información no lingüística con fines de análisis automático.

Una aplicación interesante está en el diseño de sistemas automáticos que caracterizan la presencia y gravedad de desordenes de voz. Lo cual tiene aplicaciones como herramientas complementarias objetivas en entornos clínicos. No obstante, el diseño de sistemas automáticos plantea varios problemas que incluyen la variabilidad intrínseca del habla, la presencia simultánea de múltiples fenómenos de patología vocal, información extralingüística espuria o la dependencia en evaluaciones perceptuales altamente subjetivas.

Con estos antecedentes, esta tesis evalúa la influencia de la información extralingüística, diferentes tipos de tareas de producción de habla, diversas máquinas de decisión y características, en el diseño de *sistemas automáticos de análisis de calidad vocal*, cuyo objetivo es generalizar decisiones acerca de la presencia y severidad de patologías presentes en la voz y/o el habla. Una nueva metodología ha sido propuesta para emular las capacidades perceptuales de un evaluador humano, la cual está basada en algoritmos de selección de características, clasificación ordinal y regresión gaussiana. El regresor se usa para convertir la escala de percepción discreta en una continua, más acorde con la naturaleza de las evaluaciones. Además, la robustez de los sistemas es evaluada en configuraciones de bases de datos cruzadas.

Los resultados indican que el sexo del hablante juega un papel importante en los sistemas automáticos de análisis de calidad de voz y que el diseño basado en sistemas jerárquicos debe ser considerado. También se ha encontrado que el conjunto más consistente de características en tareas de detección y evaluación de patologías son dos medidas de perturbación y un descriptor basado en la dispersión de las representaciones de *espectros de modulación: glottal-to-noise excitation ratio, cepstral harmonics-to-noise ratio y rate of points above linear average*. El mejor detector automático entrenado con la base de datos de *Saarbrücken* logra un AUC de 0.88 cuando la información provista por las diferentes tareas de voz se fusiona mediante regresión logística. En escenarios de bases de datos cruzadas, el AUC varía entre 0.75 y 0.94, lo que demuestra la solidez del sistema. Este valor constituye una de las mejores eficiencias reportadas usando esta partición. El mejor sistema de evaluación incurre en errores que difieren, en promedio, en media unidad con respecto a la etiqueta real en configuraciones de bases de datos cruzadas, usando $\mathcal{G}$ y $\mathcal{B}$. Su capacidad de generalizar resultados ha sido validada por un experto. El error del sistema evaluado clínicamente es de 0.3 unidades para $\mathcal{G}$.

PUBLICATIONS

Journal papers

- [1] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, *et al.*, «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.
- [2] J. I. Godino-Llorente, S. Shattuck-Hufnagel, J. Choi, L. Moro-Velázquez, and J. A. Gómez-García, «Towards the identification of idiopathic parkinson disease from speech. new articulatory kinetic biomarkers,» *PLOS ONE*, vol. 12, no. 12, pp. 1–35, 2017.
- [3] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, «Voice pathology detection using modulation spectrum optimized metrics,» *Frontiers in Bioengineering and Biotechnology*, vol. 4, p. 1, 2016.
- [4] J. A. Gómez-García, L. Moro-Velázquez, J. I. Godino-Llorente, and G. Castellanos Domínguez, «An insight to the automatic categorization of speakers according to sex and its application to the detection of voice pathologies: A comparative study,» *Revista Universidad de Antioquia.*, no. 79, pp. 50–62, 2016.
- [5] L. Moro-Velázquez, J. A. Gómez-García, 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. Article ID: 259239, 2015.
- [6] G. Andrade-Miranda, J. I. Godino-Llorente, L. Moro-Velázquez, and J. A. Gómez-García, «An automatic method to detect and track the glottal gap from high speed videoendoscopic images,» *Biomedical engineering online*, vol. 14, no. 1, p. 100, 2015.
- [7] J. A. Gómez-García, J. I. Godino-Llorente, and G. Castellanos-Domínguez, «Non uniform Embedding based on Relevance Analysis with reduced computational complexity: Application to the detection of pathologies from bio-signal recordings,» *Neurocomputing*, vol. 132, pp. 148–158, 2014.

Conference papers

- [1] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, «On the design of a voice pathology assessment system based on the grbas scale,» in *10th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2017.
- [2] L. Moro-Velázquez, J. I. Godino-Llorente, J. A. Gómez-García, J. Villalba, and N. Dehak, «Use of acoustic landmarks and gmm-ubm blend in the automatic detection of parkinson's disease,» in *10th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2017.
- [3] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, «Detection of parkinson's disease by means of GMM-UBM and i-vectors techniques,» in *XXXI Simposium Nacional de la Unión Científica Internacional de Radio*, 2016.

- [4] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, «Tuning of modulation spectrum parameters for voice pathology detection,» in *9th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2015.
- [5] J. A. Gómez-García, L. Moro-Velázquez, J. I. Godino-Llorente, and G. Castellanos Domínguez, «Automatic age detection in normal and pathological voice,» in *16th Annual Conference of the International Speech Communication Association. INTERSPEECH*, 2015.
- [6] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, «Analysis of complexity and modulation spectra parameterizations to characterize voice roughness,» in *VIII Jornadas en Tecnologías del Habla. IberSPEECH*, 2014.
- [7] J. A. Gómez-García, J. Blanco-Murillo, J. I. Godino-Llorente, L. Hernández-Gómez, and G. Castellanos-Domínguez, «GMM-based classifiers for the automatic detection of obstructive sleep apnea,» in *6th International Joint Conference on Biomedical Engineering Systems and Technologies*, 2013, pp. 364–367.
- [8] J. A. Gómez-García, J. I. Godino-Llorente, and G. Castellanos-Domínguez, «Automatic gender recognition in normal and pathological speech,» in *14th Annual Conference of the International Speech Communication Association. INTERSPEECH*, 2013, pp. 1707–1711.
- [9] J. A. Gómez-García, J. I. Godino-Llorente, and G. Castellanos-Domínguez, «Identificación de género para la detección automática de patologías,» in *I Jornadas Multidisciplinares de Usuarios de la voz, el habla y el canto*, 2013.
- [10] J. A. Gómez-García, J. I. Godino-Llorente, and G. Castellanos-Domínguez, «Sex-dependent automatic detection of voice pathologies,» in *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2013.
- [11] J. A. Gómez-García, J. I. Godino-Llorente, and G. Castellanos-Domínguez, «Influence of delay time on regularity estimation for voice pathology detection,» in *Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE*, 2012, pp. 4217–4220.
- [12] J. A. Gómez-García, J. I. Godino-Llorente, and G. Castellanos-Domínguez, «Speaker recognition techniques employed for pathological voice detection,» in *Workshop tecnologías multibiométricas para la identificación de personas*, 2012.
- [13] J. A. Gómez-García, J. I. Godino Llorente, and G. Castellanos-Domínguez, «Complexity analysis using nonuniform embedding techniques for pathological voice discrimination,» in *NOn Llinear Speech Processing Conference*, 2011.

Submitted papers/Work in progress

- [1] J. A. Gómez-García, S. Raimondo, P. Van-Lieshout, J. I. Godino-Llorente, and F. Rudzicz, «Task dynamics for the analysis of dysarthric patients,» *Work in progress*, 2018.
- [2] J. A. Gómez-García, L. Moro-Velázquez, J. Mendes-Laureano, C.-D. G., and J. I. Godino-Llorente, «A machine learning approach to emulate the perceptual capabilities of a human evaluator mapping the grb scale,» *Submitted-Pattern Recognition Journal*, 2018.
- [3] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, «On the effects of variability in the design of automatic detectors of voice pathologies,» *Work in progress*, 2018.

CONTENTS

I INTRODUCTION

1	SPEECH & SPEECH PRODUCTION	3
1.1	The source-filter model of speech production	4
1.2	Information contained in speech	7
1.3	Normal and pathological speech	8
1.4	Discussion	9
2	PROBLEM STATEMENT	11
3	OBJECTIVES	13

II VOICE PATHOLOGY & AUTOMATIC VOICE QUALITY ANALYSIS

4	VOICE DISORDERS	17
4.1	Categorization of voice disorders	18
4.2	Perceptual categorization of voice disorders	19
4.2.1	Disorders affecting pitch	19
4.2.2	Disorders affecting loudness	19
4.2.3	Disorders affecting quality	20
4.2.4	Disorders affecting variability	21
4.3	Discussion	22
5	AUTOMATIC VOICE QUALITY ANALYSIS SYSTEMS	23
5.1	Speech production tasks	24
5.1.1	Sustained phonation:	24
5.1.2	Running Speech	25
5.2	Decision tasks	26
5.2.1	Voice pathology detection and identification:	26
5.2.2	Voice pathology assessment	26
5.3	Discussion	29
6	VARIABILITY ASPECTS AFFECTING VOICE QUALITY ANALYSIS SYSTEMS	31
6.1	Intra-class variability	31
6.1.1	Dialects and accents	32
6.1.2	Vocal effort and loudness	32
6.1.3	Emotion	33
6.1.4	Sex	34
6.1.5	Age	35
6.2	Channel-dependent external influences	36
6.3	Discussion	36

III DESIGN OF AUTOMATIC VOICE QUALITY ANALYSIS SYSTEMS

7	STATE OF THE ART	41
7.1	Databases	41
7.2	Pre-processing	42

7.3	Characterisation	43
7.3.1	Temporal and acoustical analysis	43
7.3.2	Perturbation and fluctuation analysis	43
7.3.3	Spectral and cepstral analysis	44
7.3.4	Complexity features	45
7.3.5	3-dimensional representations	46
7.3.6	Other types of features	46
7.4	Dimensionality reduction	47
7.5	Machine learning and decision making	47
7.6	Some relevant works in Automatic Voice Quality Analysis systems	48
7.6.1	Temporal-Perturbation	48
7.6.2	Spectral and cepstral analysis	49
7.6.3	Complexity	52
7.6.4	Multiples sets of features	53
7.7	Discussion	54
8	EXTRACTION OF RELEVANT PARAMETERS	57
8.1	Preprocessing	57
8.2	Characterisation	58
8.2.1	Perturbations parameters	59
8.2.2	Spectral-Cepstral analysis of speech	63
8.2.3	Modulation Spectra features	68
8.2.4	Complexity analysis	71
8.3	Dimensionality reduction	83
8.3.1	Variable ranking through correlation coefficients	84
8.3.2	Filter subset selection	86
8.4	Discussion	87
9	MACHINE LEARNING AND EVALUATION OF RESULTS	89
9.1	Classification	89
9.1.1	Gaussian Mixture Models	90
9.1.2	GMM-based factorial models	91
9.1.3	Support Vector machines	95
9.1.4	Decision making	96
9.1.5	Fusion of scores	97
9.2	Ordinal classification	97
9.2.1	Proportional Odd model	98
9.2.2	Extreme learning machines with ordered partition.	98
9.2.3	Gaussian mixture regression	99
9.3	Evaluation of results	100
9.3.1	Metrics derived from the performance of the system	101
9.3.2	Performance curves	102
9.3.3	Assessing the reliability of perceptual evaluations	103
9.4	Discussion	104
IV EXPERIMENTAL SETUP		
10	EXPERIMENTS	107
10.1	Acoustic material	107

10.1.1	<i>Hospital Universitario Príncipe de Asturias database</i>	108
10.1.2	<i>Saarbrücken Voice Disorders database</i>	108
10.1.3	<i>Hospital Gregorio Marañón database</i>	108
10.1.4	<i>EUR0M database</i>	109
10.1.5	<i>Albayzin database</i>	110
10.1.6	<i>PhoneDat database</i>	111
10.1.7	<i>Massachusetts Ear and Eye Infirmary database</i>	112
10.1.8	<i>Hospital Doctor Negrín database</i>	112
10.1.9	<i>Aplicación de las Tecnologías de la Información y Comunicaciones database</i>	113
10.2	Experiments	113
10.2.1	Detection experiments	114
10.2.2	Assessment experiments	125
10.3	Discussion	130
11	RESULTS ON VOICE PATHOLOGY DETECTION	133
11.1	Sub-test D1	133
11.1.1	Trial D1_HUPA_sex	134
11.1.2	Trial D1_SVD_sex	134
11.1.3	Trial D1_GMar_sex	134
11.1.4	Trial D1_HUPA_age	135
11.1.5	Discussions on sub-test D1	136
11.2	Sub-test D2	138
11.2.1	Trial D2_HUPA-A	138
11.2.2	Trial D2_SVD-A	140
11.2.3	Trial D2_GMar-A	141
11.2.4	Trial D2_SVD-I	142
11.2.5	Trial D2_GMar-I	143
11.2.6	Trial D2_SVD-U	144
11.2.7	Trial D2_GMar-U	146
11.2.8	Trial D2_SVD-RS	147
11.2.9	Trial D2_SVD-RS-Vd	147
11.2.10	Discussions on sub-test D2	148
11.3	Sub-test D3	151
11.3.1	Trial D3_SusPho	151
11.3.2	Trial D3_RS	152
11.3.3	Discussions on sub-test D3	154
11.4	Sub-test D4	155
11.4.1	Trial D4_SVD	157
11.4.2	Trial D4_SVD_CrossDatabase	158
11.4.3	Discussions on sub-test D4	158
11.5	General Discussion	159
12	RESULTS ON VOICE PATHOLOGY ASSESSMENT	163
12.1	Sub-test A1	163
12.1.1	Trial A1_HUPA	164
12.1.2	Trial A1_SVD	165
12.1.3	Trial A1_GMar	167
12.1.4	Discussions on sub-test A1	169

12.2	Sub-test A2	172
12.2.1	Trial A2_SVD	175
12.2.2	Trial A2_DN	180
12.2.3	Trial A2_ATIC	182
12.2.4	Discussions on sub-test A2	185
12.3	Sub-test A3	188
12.3.1	Trial A3_SVD	188
12.3.2	Trial A3_DN	189
12.3.3	Trial A3_ATIC	189
12.3.4	Discussions on sub-test A3	189
12.4	Sub-test A4	190
12.4.1	Discussions on sub-test A4	191
12.5	General discussion	194
 V CONCLUSIONS AND FUTURE WORK		
13	CONCLUSIONS AND FUTURE WORK	201
13.1	Conclusions	201
13.2	Contributions	204
13.3	Future work	205
 VI APPENDICES		
A	APPENDIX I: CORRELATION ANALYSIS FOR DIFFERENT TYPES OF SPEECH MATERIAL.	209
B	APPENDIX II: FEATURE RANKING OF THE ASSESSMENT EXPER- IMENTS.	213
REFERENCES		217

LIST OF FIGURES

Figure 1.1	Axial view of the human larynx.	4
Figure 1.2	Glottal waveform resulting from the vibration of the vocal folds	4
Figure 1.3	Sagittal representation of the vocal tract.	5
Figure 1.4	Source-filter model of speech production	5
Figure 1.5	Simplified loss-less tube model of the vocal tract. . . .	6
Figure 1.6	Spectrum and spectral envelope of a vowel /a/ calculated using a cepstrum transformation.	7
Figure 4.1	Reinke's oedema.	20
Figure 4.2	Paresis.	20
Figure 4.3	Nodules and polyps.	21
Figure 4.4	Cleft lip.	22
Figure 5.1	Depiction of a typical AVQA system.	24
Figure 5.2	Example of the VHI test.	28
Figure 5.3	Hoarseness diagram for a normophonic and a patho- logical voice.	28
Figure 8.1	Taxonomy of voice signals.	58
Figure 8.2	Algorithm for the calculation of the CHNR	61
Figure 8.3	Magnitude and noise spectrum for a normophonic and pathological voice using CHNR	61
Figure 8.4	Algorithm for the calculation of NNE	62
Figure 8.5	Magnitude and noise spectrum of a normophonic and pathological voice using NNE	62
Figure 8.6	Methodology to estimate the GNE	63
Figure 8.7	Correlation matrix during the GNE computation for a normophonic and dysphonic voice.	63
Figure 8.8	Depiction of the CPPS calculation process	65
Figure 8.9	CPPS for a normophonic and a pathological voice. . . .	65
Figure 8.10	Depiction of the mel-filterbank	67
Figure 8.11	Depiction of the MFCC calculation process	67
Figure 8.12	MFCC of a normophonic and a pathological signal. . . .	67
Figure 8.13	Depiction of the PLP calculation process	68
Figure 8.14	Depiction of the Bark filterbank	68
Figure 8.15	PLP of a normophonic and a pathological signal. . . .	68
Figure 8.16	Stages followed to compute the MS matrix.	69
Figure 8.17	Stages followed to compute the MS features.	69
Figure 8.18	MS modulus of a normophonic and dysphonic voice. .	70
Figure 8.19	CIL for a normophonic and dysphonic voice.	71
Figure 8.20	Points above $ E $ for the computation of RALA	71
Figure 8.21	Attractor reconstruction using non-uniform embedding. 73	
Figure 8.22	Estimation of D₂	77

Figure 8.23	Estimation of LLE	77
Figure 8.24	Recurrence time histograms in a normophonic and pathological voice for the computation of RPDE	78
Figure 8.25	Estimation of h^e	79
Figure 8.26	DFA for a normophonic and a pathological voice.	80
Figure 8.27	Regularity estimators for a normophonic and a pathological voice.	81
Figure 8.28	PE entropy of a normophonic and a pathological voice.	82
Figure 8.29	Computation of rHMMEn and sHMMEn for a normophonic and dysphonic voice.	83
Figure 9.1	GMM-UBM as a MAP -adaptation of UBM models.	92
Figure 9.2	IV as a transformation of a series of utterances to a single vector.	94
Figure 9.3	Example of an ELM network.	98
Figure 9.4	Typical DET curve	103
Figure 10.1	Statistics of the HUPA database.	109
Figure 10.2	Statistics of the SVD database.	110
Figure 10.3	Histogram summarising some statistics of the GMar database.	111
Figure 10.4	Statistics of the EUROM database.	111
Figure 10.5	Statistics of the DN database.	112
Figure 10.6	Statistics of the ATIC database.	113
Figure 10.7	<i>Sub-test D1</i> : hierarchical system based on decomposing the partition of speakers according to sex and age.	115
Figure 10.8	<i>Sub-test D1</i> : age distribution for the HUPA partition.	116
Figure 10.9	<i>Sub-test D1</i> : hierarchical detection of voice pathologies.	117
Figure 10.10	<i>Sub-test D2</i> : automatic detection of voice pathologies using diverse sets of features and acoustic material.	121
Figure 10.11	<i>Sub-test D2</i> : histogram of the values of d_e using utterances of the sustained vowel /a/ of the SVD , HUPA and GMar databases.	121
Figure 10.12	<i>Sub-test D3</i> : automatic detection of voice pathologies using UBM -based classifiers.	123
Figure 10.13	<i>Sub-test D4</i> : methodology of the best performing system.	125
Figure 10.14	<i>Sub-test D4</i> : scoring procedure for selecting the most consistent characteristics.	126
Figure 10.15	<i>Sub-test A1</i> : methodology for predicting $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$	127
Figure 10.16	<i>Sub-test A2</i> : automatic grading system based on the GRBAS scale.	128
Figure 10.17	<i>Sub-test A4</i> : Matlab [®] application for the blind assessment of voice quality.	130
Figure 11.1	<i>Trial D1_HUPA_sex</i> : DET curve using the HUPA database.	134
Figure 11.2	<i>Trial D1_SVD_sex</i> : DET curve using the SVD database.	135
Figure 11.3	<i>Trial D1_GMar_sex</i> : DET curve using the GMar database.	135
Figure 11.4	<i>Trial D1_HUPA_age</i> : DET curve using the HUPA database.	136

Figure 11.5	<i>Sub-test D1</i> : effects of considering age and sex in an AVQA system using the vowel /a/ and the SVD dataset.	138
Figure 11.6	<i>Trial D2_HUPA-A</i> : DET curves for different sets of features using the HUPA database and the vowel /a/.	139
Figure 11.7	<i>Trial D2_SVD-A</i> : DET Curves for different sets of features using the SVD database and the vowel /a/.	140
Figure 11.8	<i>Trial D2_GMar-A</i> : DET curves for the best set of features using the GMar database and vowel /a/.	141
Figure 11.9	<i>Trial D2_SVD-I</i> : DET curves for the best set of features using the SVD database and vowel /i/.	142
Figure 11.10	<i>Trial D2_GMar-I</i> : DET curves for the best set of features using the GMar database and vowel /i/.	144
Figure 11.11	<i>Trial D2_SVD-U</i> : DET curves for the best set of features using the SVD database and vowel /u/.	145
Figure 11.12	<i>Trial D2_GMar-U</i> : DET Curves for the best set of features using the GMar database and vowel /u/.	146
Figure 11.13	<i>Trial D2_SVD-RS</i> : DET curve of the best results.	147
Figure 11.14	<i>Trial D2_SVD-RS-Vd</i> : DET curve of the best results.	148
Figure 11.15	<i>Sub-test D2</i> : best DET curves using speech material of different nature	150
Figure 11.16	<i>Trial D3_SusPho</i> : DET curve of the best UBM -based classifiers.	152
Figure 11.17	<i>Trial D3_RS</i> : DET curve of the best UBM -based classifiers.	154
Figure 11.18	<i>Sub-test D4</i> : resulting methodology after having considered the outcomes of <i>sub-test D1</i> , <i>sub-test D2</i> and <i>sub-test D3</i>	157
Figure 11.19	<i>Trial D4_SVD</i> : DET curve for the best results in SVD	157
Figure 12.1	<i>Sub-test A2</i> : contour formed after plotting the CPPS of two voices of the HUPA dataset.	174
Figure 12.2	<i>Sub-test A2</i> : procedure for finding a compact representation of the feature space.	174
Figure 12.3	<i>Sub-test A2</i> : representation in a per-file basis of the most relevant features according to the feature ranking procedure.	175
Figure 12.4	<i>Sub-test A2</i> : pmf for the labels 0, 1, 2 or 3.	176
Figure 12.5	<i>Trial A2_SVD</i> : confusion matrices for the ordinal classification using the SVD partition.	177
Figure 12.6	<i>Trial A2_SVD</i> : weighted- AMAE for the SVD database.	178
Figure 12.7	<i>Trial A2_SVD</i> : raster plots and pdf modelling the differences between the target label and the label predicted by the system.	179
Figure 12.8	<i>Trial A2_DN</i> : confusion matrices for the ordinal classification using the DN partition.	181
Figure 12.9	<i>Trial A2_DN</i> : weighted- AMAE for the DN database.	182
Figure 12.10	<i>Trial A2_DN</i> : raster plots and pdf modelling the differences between the target label and the label predicted by the system.	183

Figure 12.11	<i>Trial A2_ATIC</i> : confusion matrices for the ordinal classification using the ATIC partition.	184
Figure 12.12	<i>Trial A2_ATIC</i> : weighted- AMAE for the ATIC database. .	185
Figure 12.13	<i>Trial A2_ATIC</i> : raster plots and pdf modelling the differences between the target label and the label predicted by the system.	186
Figure 12.14	<i>Sub-test A4</i> : raster plots and pdf of the deviation between the evaluations predicted by the system in <i>sub-test A2</i> and the corrected labels provided by the speech therapist for the SVD database.	191
Figure 12.15	<i>Sub-test A4</i> : raster plots and pdf of the deviation between the evaluations predicted by the system in <i>sub-test A2</i> and the corrected labels provided by the speech therapist for the DN database.	191
Figure 12.16	<i>Sub-test A4</i> : raster plots and pdf of the deviation between the evaluations predicted by the system in <i>sub-test A2</i> and the corrected labels provided by the speech therapist for the ATIC database.	192

LIST OF TABLES

Table 1.1	Dimensions of speech.	8
Table 8.1	Features described in this thesis for the characterisation stage in AVQA systems.	88
Table 9.1	Summary of Factorial models	95
Table 9.2	Coding matrix for the ELMOP computation.	99
Table 9.3	Depiction of a confusion matrix	101
Table 10.1	Sets and subsets of features employed during this thesis.	118
Table 10.2	Summary of all the trials to be performed in this thesis.	131
Table 10.3	Summary of all the databases employed in this thesis. .	131
Table 11.1	<i>Trial D1_HUPA_sex</i> : performance of the sex-dependent and sex-independent system using the HUPA database. .	134
Table 11.2	<i>Trial D1_SVD_sex</i> : performance of the sex-dependent and sex-independent system using the SVD database. .	135
Table 11.3	<i>Trial D1_GMar_sex</i> : performance of the sex-dependent and sex-independent system using the GMar database. .	135
Table 11.4	<i>Trial D1_HUPA_age</i> : performance of the age-dependent and age-independent system using the HUPA database. .	136
Table 11.5	<i>Trial D2_HUPA-A</i> : performance in detection tasks for all the tested sets of features using the HUPA database and the vowel /a/.	139

Table 11.6	<i>Trial D2_SVD-A</i> : performance in detection tasks for different sets of features, using the SVD database and the vowel /a/ as acoustic material.	140
Table 11.7	<i>Trial D2_GMar-A</i> : performance in detection tasks for different sets of features, using the GMar database and the vowel /a/.	142
Table 11.8	<i>Trial D2_SVD-I</i> : performance in detection tasks for different sets of features using the SVD database and the vowel /i/.	143
Table 11.9	<i>Trial D2_GMar-I</i> : performance in detection tasks for different sets of features using the GMar database and the vowel /i/.	144
Table 11.10	<i>Trial D2_SVD-U</i> : performance in detection tasks for different sets of features using the SVD database and the vowel /u/.	145
Table 11.11	<i>Trial D2_GMar-U</i> : performance in detection tasks for different sets of features using the GMar database and the vowel /u/.	146
Table 11.12	<i>Trial D2_SVD-RS</i> : best operation points using the sentences in the SVD database, for the spectral/cepstral features.	147
Table 11.13	<i>Trial D2_SVD-RS-Vd</i> : best results using voiced segments of the sentences in the SVD database.	148
Table 11.14	<i>Trial D3_SusPho</i> : performance of the UBM -based classifiers using the vowel /a/ in the SVD dataset.	152
Table 11.15	<i>Trial D3_RS</i> : performance of the UBM -based classifiers using the running speech sentences in the SVD dataset.	153
Table 11.16	<i>Sub-test D4</i> : top-ranked features from each database after having applied the scoring procedure.	156
Table 11.17	<i>Sub-test D4</i> : top-10 most consistent and generalist features according to the feature ranking procedure.	156
Table 11.18	<i>Trial D4_SVD</i> : best results using all the acoustic material in SVD	157
Table 11.19	<i>Trial D4_SVD_CrossDatabase</i> : results after using the methodology presented in <i>trial D4_SVD</i> in a cross-database scenario.	158
Table 12.1	<i>Sub-test A1</i> : correlation between pairs of traits for the different databases.	164
Table 12.2	<i>Trial A1_HUPA</i> : correlation of the tested features and $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ using the vowel /a/ in the HUPA dataset.	165
Table 12.3	<i>Trial A1_HUPA</i> : ordinal classification for the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits using the vowel /a/ in the HUPA dataset.	166
Table 12.4	<i>Trial A1_SVD</i> : correlation of the tested features and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits using the vowel /a/ of the SVD dataset.	167
Table 12.5	<i>Trial A1_SVD A1</i> : ordinal classification of $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ using the vowel /a/ in the SVD dataset.	168

Table 12.6	<i>Trial A1_GMar</i> : correlation of the tested features for $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$, using the vowel /a/ in the <i>GMar</i> dataset.	169
Table 12.7	<i>Trial A1_GMar</i> : ordinal classification for $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$, using the vowel /a/ in the <i>GMar</i> dataset.	170
Table 12.8	<i>Subtest A1</i> : best characteristics within each feature set and among all the databases, according to the filter feature selection procedures.	171
Table 12.9	<i>Sub-test A1</i> : $\rho_{\mathcal{R}}$ for U_e and NU_e for the three tested databases and the best performing features according to the feature selection procedure.	172
Table 12.10	<i>Sub-test A2</i> : top-10 ranked features, according to the scoring procedure.	173
Table 12.11	<i>Trial A2_SVD</i> : best results obtained using the ordinal classification techniques and the <i>SVD</i> testing partition.	176
Table 12.12	<i>Trial A2_SVD</i> : error measures for the Gaussian regressor.	178
Table 12.13	<i>Trial A2_DN</i> : best results obtained using the ordinal classification techniques and the <i>DN</i> partition.	180
Table 12.14	<i>Trial A2_DN</i> : error measures for the Gaussian regressor.	182
Table 12.15	<i>Trial A2_ATIC</i> : best results obtained using the ordinal classification techniques and the <i>ATIC</i> partition.	182
Table 12.16	<i>Trial A2_ATIC</i> : error measures for the Gaussian regressor.	185
Table 12.17	<i>Trial A3_SVD</i> : best results of the sex-dependent system obtained using the ordinal classification techniques and the <i>SVD</i> testing partition.	188
Table 12.18	<i>Trial A3_DN</i> : best results of the sex-dependent system obtained using the ordinal classification techniques and the <i>DN</i> testing partition.	189
Table 12.19	<i>Trial A3_ATIC</i> : best results of the sex-dependent system obtained using the ordinal classification techniques and the <i>ATIC</i> testing partition.	189
Table 12.20	<i>Sub-test A3</i> : comparison of the results of the sex-dependent (S.D.) and sex-independent (S.I.) assessment systems in terms of <i>Average Mean Absolute Error (AMAE)</i>	190
Table 12.21	<i>Sub-test A4</i> : error measures between the labels predicted by the automatic system and the corrections made by the speech therapist.	191
Table 12.22	Databases used for evaluating the consistency of the expert. The number and type of assessments for each database is presented.	192
Table 12.23	<i>Sub-test A4</i> : consistency between continuous assessments of the reduced <i>GRBAS</i> scale for the <i>SVD</i> partition and <i>DN</i> database.	192
Table 12.24	<i>Sub-test A4</i> : consistency between assessments using the discrete and continue evaluations of the reduced <i>GRBAS</i> scale.	193

Table 12.25	<i>Sub-test A4</i> : consistency between the evaluations provided by the speech therapist and three experts in consensus for the ATIC corpus.	193
Table A.1	<i>Trial D2_HUPA-A</i> : correlation analysis between all the tested characteristics and their labels, using the HUPA database and the vowel /a/.	209
Table A.2	<i>Trial D2_SVD-A</i> : correlation analysis between all the tested characteristics and their labels using the SVD database and the vowel /a/.	210
Table A.3	<i>Trial D2_GMar-A</i> : correlation analysis between all the tested characteristics and their labels using the GMar database and the vowel /a/.	210
Table A.4	<i>Trial D2_SVD-I</i> : correlation analysis between all the tested characteristics and their labels using the SVD database and the vowel /i/.	211
Table A.5	<i>Trial D2_GMar-I</i> : correlation analysis between all the tested characteristics and their labels using the GMar database and the vowel /i/.	211
Table A.6	<i>Trial D2_SVD-U</i> : correlation analysis between all the tested characteristics and their labels using the SVD database and the vowel /u/.	212
Table A.7	<i>Trial D2_GMar-U</i> : correlation analysis between all the tested characteristics and their labels using the GMar database and the vowel /u/.	212
Table B.1	<i>Sub-test A1</i> : best features ranked according to the feature selection algorithms for the vowel /a/ in HUPA , for $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$	213
Table B.2	<i>Sub-test A1</i> : best features ranked according to the feature selection algorithms for the vowel /a/ in SVD , for $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$	214
Table B.3	<i>Sub-test A1</i> : best features ranked according to the feature selection algorithms for the vowel /a/ in GMar , for $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$	215

NOTATION

$s[\cdot]$	Speech signal	$\{s[t]\}_{t=1}^T$
$\vec{x}[\cdot]$	Vector of features	$\{x[i]\}_{i=1}^d$
$u[\cdot]$	Excitation source in the source-filter model of speech	
$h[\cdot]$	Filter in the source-filter model of speech	
f_0	Fundamental frequency	
$f_1, f_2, \dots$	Formant frequencies	
$w[\cdot]$	Tapering window function	Length W
$s_f[\cdot]$	Windowed speech signal	$\{s_f[t]\}_{f=1}^F$
$\mathcal{F}\{\cdot\}$	Fourier transform	
$\mathbb{S}[k]$	Speech spectrum	Order K
$\mathfrak{C}\{\cdot\}$	Cepstrum transform	
ι_{En}	Harmonics energy	
γ_{En}	Noise energy	
$F^{Hz}, F^{Bark}, F^{mel}$	Frequency expressed in Hz, Bark or mel respectively	
$\mathbf{E}$	MS matrix	$A \times M$
$\mathbf{E}(f^a, f^m)$	Point of the MS matrix at acoustic frequency f^a and modulation frequency f^m	
$\vec{s}[t]$	Embedded stated vector	Dimension d_e
d_e	Embedding dimension	Scalar
d_w	Embedding window	Scalar
τ	Time lag	Scalar
$\vec{\zeta}$	Vector of non-uniform time lags	
$\mathbf{M}$	Expanded embedding matrix	$\mathbf{M} \in \mathbb{R}^{(T-d_w^{max}) \times d_w^{max}}$
$\vec{\eta}^i$	i -th column associated to a matrix	
$\mathcal{H}(\cdot)$	Entropy operator	
$\mathcal{I}(\cdot, \cdot)$	Mutual information operator	
$\mathfrak{d}(\vec{s}[t_i], \vec{s}[t_j]; \epsilon)$	Distance function between two state vectors within a radio ϵ	

$\vartheta(\cdot)$	Heaviside function	
$\mathfrak{I}(\cdot)$	Taken's operator	
$\mathcal{U}(\vec{s}[t_i]; \epsilon)$	Neighbourhood of state vector $\vec{s}[t_i]$ enclosed by a ball of radius ϵ	
$p(\Delta t_r)$	Recurrence-time probability density	
$\mathfrak{B}(\cdot)$	Auxiliary function for the computation of DFA	
$\phi^{d_m}(r); \mathring{\phi}^{d_m}(r)$	Auxiliary function computation of regularity	
r	Tolerance value for the regularity estimators	
ρ_F	Parameter defining the shape of the fuzzy function in FuzzyEn	
d_m	Dimension of state vectors for the computation of regularity	
$\vec{s}_{\pi_i}$	Permutation pattern	
$p(\vec{s}_{\pi_i})$	Permutation probability	
$\mathfrak{S}_t$	State of DHMM process at time t	
n_θ	number of states in the Markov chain	
σ	Initial state distribution in DHMM	
$\mathfrak{P}$	set of transition probabilities among states	
$\mathfrak{D}$	probability distribution of the observation symbol	
n_v	Total number of symbols	
r_α	Entropy order	Scalar
$\mathcal{X}$	Collection of training data vectors $\vec{x}$	$\{\vec{x}_n\}_{n=1}^N$
$\vec{\ell}$	Vector of labels associated to $\mathcal{X}$	$\{\ell_n\}_{n=1}^N$
$\mathcal{N}$	Matrix of characteristics	$\mathcal{N} \in \mathbb{R}^{N \times d}$
$\mathring{\mathcal{N}}$	Optimised matrix of characteristics	$\mathring{\mathcal{N}} \in \mathbb{R}^{N \times \hat{d}}$
$\rho(\cdot, \cdot)$	Pearson correlation	
$\text{cov}(\cdot, \cdot)$	Covariance operator	

$\text{std}(\cdot)$	Standard deviation operator	
$\text{var}(\cdot)$	Variance operator	
$\rho_{\mathcal{R}}$	Distance correlation	
k	Relevance index	
$\mathcal{N}(\cdot)$	Gaussian probability density function	
$p(\cdot)$	Probability density function	
λ	Mixture weight	scalar
$\vec{\mu}$	Mean of Gaussian pdf	Dimension $\vec{\mu} \in \mathbb{R}^d$
Σ	Covariance of Gaussian pdf	$\Sigma \in \mathbb{R}^{d \times d}$
Θ	Set of parameters composed of $\{\vec{\lambda}, \vec{\mu}, \Sigma\}$	
$\mathcal{F}$	First order sufficient statistic	
$\mathcal{Z}$	Zero-th order sufficient statistic	
$\mathcal{R}$	Responsibilities	
α	MAP adaptation coefficient	
β	MAP relevance factor	
$\psi(\cdot)$	Transformation from data vector to supervector	
$\vec{m}$	Supervector	$\vec{m} \in \mathbb{R}^{G \cdot d}$
$\vec{m}^u$	Supervector referred to the UBM model	
$\vec{d}_n^o$	Class and recording-specific offset	
D	Diagonal covariance matrix of residuals	$D \in \mathbb{R}^{G \cdot d \times G \cdot d}$
$\vec{z}$	Latent residual vector	$\vec{z} \in \mathbb{R}^{G \cdot d \times 1}$
V	Eigenvoice matrix	$V \in \mathbb{R}^{G \cdot d \times r_v}$
r_v	Number of eigenvoices	
$\vec{y}$	Latent class vector	$\vec{y} \in \mathbb{R}^{r_v \times 1}$
U	Intraclass-varitions matrix	$U \in \mathbb{R}^{G \cdot d \times r_u}$
r_u	Dimension intraclass-varitions matrix	
$\vec{x}$	Latent intraclass vector	$\vec{x} \in \mathbb{R}^{r_u \times 1}$
T	Total variability matrix	$T \in \mathbb{R}^{G \cdot d \times r_v}$
$\vec{w}$	i -vector	$\vec{w} \in \mathbb{R}^{r_t \times 1}$
r_t	Dimension i -vector	
Φ	Eigenvoice matrix PLDA decomposition	$\Phi \in \mathbb{R}^{r_t \times r_h}$

$\vec{v}$	Latent vector PLDA decomposition	$\vec{v} \in \mathbb{R}^{r_h \times 1}$
r_h	Dimension PLDA latent-vector	
$\mathbf{W}$	Whitening matrix	
c_l	Ideal outputs SVM model	
ξ_l	SVM weights	
$\vec{\delta}_l$	Support vectors	
$\mathcal{K}(\cdot, \cdot)$	Kernel function	
$\Lambda(\cdot)$	Log-likelihood decision function	
f_β	f-Score	
$\prec$	Order relation operator	
$\wp$	Linear transformation POM models	
$\mathfrak{T}$	Coded label matrix ELMOP	$\{\mathfrak{T}_i\}_{i=1}^Q$
$\mathcal{M}$	Hidden layer output matrix	
$\mathcal{W}$	Weight matrix relating hidden and output layers	
$\vec{\mathcal{E}}_n$	Concatenated training sample and label, $[\vec{x}_n, \ell_n]$	
$\underline{h}(\cdot)$	Auxiliary function Gaussian regression	
$\mathcal{O}[\cdot]$	Operator indicating the position of the label in the ordinal rank	
$\mathfrak{U}$	Generalisability coefficient	

ACRONYMS

$\mathcal{A}$	<i>Asthenia</i>
ACC	<i>Accuracy</i>
AMAE	<i>Average Mean Absolute Error</i>
ANN	<i>Artificial Neural Networks</i>
ANOVA	<i>Analysis of Variance</i>
ApEn	<i>Approximate Entropy</i>
APQ ₃	<i>Shimmer Three-point Amplitude Perturbation Quotient</i>
ATIC	<i>Aplicación de las Tecnologías de la Información y Comunicaciones</i>
AUC	<i>Area Under Receiver Operating Characteristic Curve</i>
AVQA	<i>Automatic Voice Quality Analysis</i>
$\mathcal{B}$	<i>Breathiness</i>
CAPE-V	<i>Consensus Auditory-Perceptual Evaluation of Voice</i>
CHNR	<i>Cepstral Harmonics-to-Noise Ratio</i>
CIL	<i>Cumulative Intersection Point</i>
CPP	<i>Cepstral Peak Prominence</i>
CPPS	<i>Smoothed Cepstral Peak Prominence</i>
D_2	<i>Correlation Dimension</i>
DCT	<i>Discrete Cosine Transform</i>
DET	<i>Detection Error Trade-off</i>
DFA	<i>Detrended Fluctuations Analysis</i>
DHMM	<i>Discrete Hidden Markov model</i>
DN	<i>Hospital Doctor Negrín</i>
DTFT	<i>Discrete Time Fourier Transform</i>
DynInv	<i>Dynamic Invariants</i>
EER	<i>Equal Error Rate</i>
EGG	<i>Electroglotography</i>
EIG	<i>Eigenvoice Adaptation</i>
ELM	<i>Extreme Learning Machine</i>
ELMOP	<i>Extreme Learning Machine with Ordered Partitions</i>
EM	<i>Expectation-Maximization</i>

EntAtt	<i>Entropy Features based on the Reconstruction</i>
FFT	<i>Fast Fourier Transform</i>
FN	<i>False Negative</i>
FNN	<i>False Nearest Neighbours</i>
FNr	<i>False Negative Rate</i>
FP	<i>False Positive</i>
FPr	<i>False Positive Rate</i>
FuzzyEn	<i>Fuzzy Entropy</i>
$\mathcal{G}$	<i>Grade</i>
GApEn	<i>Gaussian Kernel Approximate Entropy</i>
GMar	<i>Hospital Gregorio Marañón</i>
GMM	<i>Gaussian Mixture Model</i>
GMM-UBM	<i>Adapted Model coming from an Universal Background Model</i>
GMM-SVM	<i>Gaussian Mixture Models - Support Vector Machine</i>
GMR	<i>Gaussian Mixture Regression</i>
GNE	<i>Glottal-to-Noise Excitation Ratio</i>
GRBAS	<i>Grade-Roughness-Breathiness-Asthenia-Strain</i>
GSampEn	<i>Gaussian Kernel Sample Entropy</i>
h^e	<i>Hurst Exponent</i>
HMM	<i>Hidden Markov Model</i>
HMP	<i>Hidden Markov Process</i>
HNR	<i>Harmonics-to-Noise Ratio</i>
HUPA	<i>Hospital Universitario Príncipe de Asturias</i>
ISC	<i>Intersession Variability Compensation</i>
IV	<i>i-Vector Modelling</i>
JFA	<i>Joint Factor Analysis</i>
JMI	<i>Joint Mutual Information</i>
KNN	<i>k-Nearest Neighbour</i>
LDA	<i>Linear Discriminant Analysis</i>
LHr	<i>Low-to-High Frequency Spectral Energy Ratio</i>
LLE	<i>Largest Lyapunov Exponent</i>
LongRange	<i>Long-range Correlation</i>
LPC	<i>Linear Prediction Coding Coefficients</i>
LPCC	<i>Linear Prediction Cepstral Coefficients</i>

LTAS	<i>Long-Time Average Spectrum</i>
MAE	<i>Mean Absolute Error</i>
MAP	<i>Maximum a-posterior</i>
MAUS	<i>Munich Automatic Segmentation system</i>
MC	<i>Markov Chains</i>
MDVP	<i>Multidimensional Voice Program</i>
MEEI	<i>Massachusetts Ear and Eye Infirmary</i>
MDL	<i>Minimum Description Length</i>
MFCC	<i>Mel-frequency Cepstral Coefficients</i>
MIM	<i>Mutual Information Maximisation</i>
ML	<i>Maximum Likelihood</i>
mRMR	<i>Max-Relevance Min-Redundancy</i>
MS	<i>Modulation Spectra</i>
MSs	<i>Modulation Spectrum</i>
mSampEn	<i>Modified Sample Entropy</i>
MSH	<i>Modulation Spectra Homogeneity</i>
MSP	<i>Modulation Spectrum Percentile</i>
NDA	<i>Nonlinear Dynamics Analysis</i>
NNE	<i>Normalized Noise Energy</i>
NU_e	<i>Non-Uniform</i>
PE	<i>Permutation Entropy</i>
Pert	<i>Perturbation</i>
PCA	<i>Principal Component Analysis</i>
PLDA	<i>Probabilistic Linear Discriminant Analysis</i>
PLP	<i>Perceptual Linear Prediction coefficients</i>
POM	<i>Proportional Odd Model</i>
pdf	<i>Probability Distribution Function</i>
pmf	<i>Probability Mass Function</i>
PPQ ₅	<i>Jitter Five-point Period Perturbation Quotient</i>
$\mathcal{R}$	<i>Roughness</i>
RALA	<i>Rate of Points Above Linear Average</i>
RAP	<i>Jitter Relative Average Perturbation</i>
RBH	<i>Roughness-Breathiness-Hoarseness</i>
Reg	<i>Regularity Features</i>

rHMMEn	<i>Rényi HMM Entropy</i>
RMSE	<i>Root Mean Square Error</i>
ROC	<i>Receiver Operating Characteristic</i>
RPDE	<i>Recurrence Period Density Entropy</i>
$\mathcal{S}$	<i>Strain</i>
SampEn	<i>Sample Entropy</i>
SE	<i>Sensitivity</i>
sHMMEn	<i>Shannon HMM Entropy</i>
SpecCeps	<i>Spectral/Cepstral</i>
SNR	<i>Signal-to-Noise Ratio</i>
SP	<i>Specificity</i>
SVD	<i>Saarbrücken Voice Disorders</i>
SVM	<i>Support Vector Machines</i>
sTFT	<i>Short-Time Fourier Transform</i>
TN	<i>True Negative</i>
TP	<i>True Positive</i>
TPr	<i>True positive rate</i>
UBM	<i>Universal Background Model</i>
UmRMR	<i>Unsupervised Minimal Redundancy/Maximal Relevance</i>
U_e	<i>Uniform Embedding</i>
VHI	<i>Voice Handicap Index</i>

Part I

INTRODUCTION

SPEECH & SPEECH PRODUCTION

Undoubtedly the ability to produce speech and encode meaning in the form of language, has played an important role in the advancement of society, allowing humans to exchange information in efficient and simple manners through the use of verbal means. Despite speech is produced in everyday basis, the act of speaking involves an extremely complex process resulting from the precise coordination of several subsystems acting in conjunction to generate a meaningful audible output. In this respect, the subsystems involved in the production of speech include the respiratory, phonatory, articulatory, resonant and nervous systems as described next [1–5]:

- The **respiratory** or **breathing** subsystem comprises the structures below the larynx, including the respiratory passageway, lungs, trachea, etc. It provides the driving force for speech production.
- The **phonatory** or **laryngeal** subsystem is composed of the *larynx* and is involved in the production of voiced sounds. A graphic depicting some of the cartilaginous structures of the larynx is presented in Figure 1.1a. Two multi-layered folds of tissue within the larynx called *vocal folds* or *vocal chords* serve as valves that permit or restrain the flux of air coming from the lungs. They are also caused to vibrate when the air flows through, resulting into audible vocal sounds. The space that is formed between both vocal folds is termed *glottis*, *glottal gap* or *glottal slit*. In the same manner, the pulses of air pressure resulting from opening and closing the glottis conform a *glottal volume-velocity waveform*, *glottal flow* or simply *glottal waveform*. As illustrated by Figure 1.2, the glottal waveform is mainly composed by three key phases of vibration: closed phase (vocal folds being together), opening phase (vocal folds parting), and closing phase (vocal folds coming together). Frequently, the opening and closing phases are simply referred to as open phase because this is the time during which air flows [6].
- The **articulatory** or **pharyngeal-oral** subsystem is composed by articulators that alter the characteristics of the glottal airflow coming from the lungs with structures like the tongue, lips, teeth, velum, etc.
- The **resonant** or **velopharyngeal-nasal** subsystem is in charge of adjusting the coupling between the *pharyngeal cavity* (extending from the top of trachea to the velum) and the *nasal cavity* (extending from the velum to the nostrils). During speech production, the size of the

velopharyngeal port varies depending on the nature of the speech that is produced. For instance when the velum is lowered, there is coupling to produce nasal sounds like /m/, /n/ or /ng/.

- The **nervous** subsystem is in charge of controlling the phonatory, articulatory and resonant subsystems for the production of speech. It provides the intelligence for the fine process of speech generation.

Figure 1.1. Axial view of the human larynx during the (a) voicing state; and (b) unvoicing state. The arrows indicate the direction of movement, and the fuzzy lines the presence of turbulence. Graphic from [5].

Usually, the phonatory, articulatory and resonant subsystems are grouped into a superstructure called *vocal tract* -beginning at the glottis and ending at the lips- that encloses the most important mechanical structures involved in the production of spoken sounds. A schematic of the vocal tract and some of the subsystems involved in speech production is presented in Figure 1.3.

1.1 THE SOURCE-FILTER MODEL OF SPEECH PRODUCTION

The speech generation process can be described mathematically by means of a simplified *source-filter* model, that assumes speech as the result of the convolution between an *excitation input* or *source* (conformed by the glottal waveform resulting from the vibration of the vocal folds) and a *filter* (formed by the vocal tract). A graphic depicting this simplified source-filter model is shown in Figure 1.4, where $u[t]$ is the excitation source, $h[t]$ represents the

Figure 1.2. Glottal waveform resulting from the vibration of the vocal folds. Image extracted from [6].

Figure 1.3. Sagittal representation of the vocal tract. Image adapted from [7].

filter (with transfer function $H(f)$), and $s[t]$ is the output speech signal after convolving source and filter:

$$s[t] = u[t] * h[t] \quad (1.1)$$

Figure 1.4. Source-filter model of speech production

Within this source-filter modelling framework two types of excitation inputs -corresponding to the states of the vocal folds relevant for speech production - are considered: voicing and unvoicing [5]. On one hand, during the *voicing* state, the arytenoid cartilages move towards each other as air is expelled from the lungs, the vocal folds get close together and vibrate at a rate defined by the **fundamental frequency** (f_0) in a process called *phonation*. This voicing state is typical in the production of *voiced sounds*, like in vowels /a/, /e/, On the other hand, during the *unvoicing* state, the vocal folds get close together and tense with no vibration, allowing *turbulence* to be generated. This turbulence is also known as *aspiration noise* or simply *aspiration*, and is modelled with a white-noise excitation source. The unvoicing state is typical in the production of *unvoiced sounds*, like in certain consonants, e.g. /p/, /t/, Figure 1.1 summarises the states of the arytenoid cartilages and the vocal folds during the voicing and unvoicing states.

In order to describe the filter within the source-filter model, the resonant properties of the vocal tract should be analysed first. To this end, the vocal tract is often simplified by means of a loss-less tube model that considers

the vocal apparatus as a long tube closed at the glottis and open at the lips. Figure 1.5 illustrates the simplified model in the production of a neutral vowel for an adult male whose vocal tract is 17.5 cm length and $f_0 = 340$ Hz. The frequency at which the tube resonates most of its energy receives the name of *first resonant frequency* or *first formant* (f_1), and is achieved with a wavelength 4 times the length of the tube, i. e., $340 / (4 \times 0.175) = 485.7$ Hz. Theoretically, infinite frequencies are produced at odd multiples of f_1 such that $f_2 = 3f_1$, $f_3 = 5f_1$, $\dots$, but only the first 4 or 5 are considered relevant for speech perception and production [8].

Figure 1.5. Simplified loss-less tube model of the vocal tract apparatus. Image modified from [9].

Even though the tube model works well for characterizing some spoken sounds (specially open vowels), it is an oversimplification that fails at completely representing the resonance phenomena of the vocal tract. Alternatively, the resonant frequencies of the vocal tract can be studied using the *speech spectrum*. The spectrum, describes the intensity (and phase) of the frequency components conforming the signal, and is typically obtained using the *Short-Time Fourier Transform* (**sTFT**) (through a *Fast Fourier Transform* (**FFT**) implementation). In this manner, for a certain fragment of speech $\underline{s}[t]$ of length T (e. g., a frame resulting from a windowing procedure as in Chapter 8), the **sTFT** is computed as [10]:

$$\mathcal{F}\{\underline{s}[t]\} := \{\mathbf{S}[k]\}_{k=0}^{K-1} = \sum_{t=0}^{T-1} \underline{s}[t] \exp \frac{-j2\pi tk}{K} \quad (1.2)$$

where $\mathcal{F}\{\cdot\}$ stands for the Fourier transform, $\{\mathbf{S}[\cdot]\}$ is the resulting speech spectrum and K is the number of coefficients in the transform. Frequently, the phase information of the previous expression is disregarded, leading to the computation of the power spectrum $|\mathbf{S}[\cdot]|$ (magnitude of the spectrum).

Another manner of describing the intensity of periodic patterns in the magnitude spectrum is through *cepstral analysis* of speech. The *cepstrum* is the Fourier transform of the log-power spectrum which is defined as:

$$\mathcal{C}\{\underline{s}[t]\} := \mathcal{F}^{-1}\{\log|\mathcal{F}\{\underline{s}[t]\}|\} \quad (1.3)$$

where $\mathcal{F}^{-1}\{\cdot\}$ stands for the inverse Fourier transform and $\mathcal{C}\{\cdot\}$ is the resulting cepstral transformation.

An interesting property of the cepstrum is that it converts convolution in addition. This is the so called *decorrelating property* of the cepstrum, and is often employed to decompose speech into its excitation source and vocal tract components. In this manner, Equation (1.1) can be rewritten as [11]:

$$\mathfrak{C}\{s[t]\} = \mathfrak{C}\{u[t] * h[t]\} = \mathfrak{C}\{u[t]\} + \mathfrak{C}\{h[t]\} \quad (1.4)$$

An example of the cepstrum applied to the analysis of the resonant properties of the speech is presented in Figure 1.6, on which the spectrum of a vowel /a/ is plotted and superimposed to the cepstrum-smoothed spectrum that is obtained after having removed the source component. In other words, the spectrum presents all the frequency content that results from the convolution of both source and filter, whereas the cepstrum-smoothed spectrum serves as an estimator of the *spectral envelope* of the speech signal when no excitation is considered, and therefore is related to the resonant properties of the vocal tract. Indeed, the peaks of this spectral envelope coincide with the resonant frequencies of the vocal tract.

Figure 1.6. Spectrum and spectral envelope calculated using a cepstrum transformation for a register of a vowel /a/. The peaks of the spectral envelope coincide with the formants of the vocal tract.

1.2 INFORMATION CONTAINED IN SPEECH

Despite the main objective of speech is transmitting information by means of sounds that encode linguistic content, the inherent intricacy of the speech production process embeds a substantial amount of non-linguistic aspects into speech signals. Several authors have tried to categorize the information enclosed in the speech by using different descriptors. For instance Laver [12] -in a classic definition- states that the verbal forms of speech contain a linguistic dimension associated to the use of phonological and grammatical units. A paralinguistic dimension that is communicative, non-verbal and non-linguistic, conveying information about the affective, attitudinal or emotional state of the speaker. And an extralinguistic dimension that is not communicative but which comprises information about the speaker itself. Other authors, such as Traunmuller [13, 14] have proposed a categorization using a

phonetic or linguistic dimension which is related to the message, variations in language, dialect, sociolect, idiolect and speech style of the speaker, and which is reflected by means of words, sounds, prosodic patterns, etc. An **affective, expressive or paralinguistic** dimension where the remaining communicative aspects of speech not transmitted linguistically are embodied, informing about emotions, attitudes, etc., by means of the type of phonation (modal, creaky, etc.), register, vocal effort, speech rate, etc. A **personal, organic or extralinguistic** dimension which is not communicative but informative about the speaker's identity and state, reflecting characteristics such as age, sex, pathology, etc. And a **transmittal or perspectival** dimension which tells nothing about the speaker or its message but about its physical location. Since the Traunmuller's definition accounts for factors that the one presented by Laver neglects, this is preferred through the development of this thesis. Having this in mind, Table 1.1 summarises the dimensions of speech according to above-stated Traunmuller's definition.

Dimension	Content	Manifestation
Linguistic	Message, speaking style, dialect, ...	Words, prosodic patterns, sounds, ...
Paralinguistic	Emotions, attitudes, ...	Type of phonation, vocal effort, speech rate, ...
Extralinguistic	Speaker's state, age, sex, ...	Larynx size, vocal tract, ...
Transmittal	Physical location, orientation, ...	Environmental conditions, ...

Table 1.1. Dimensions of speech based on the definition of [13, 14].

It is important to remark that among all the possible linguistic and non-linguistic dimensions which exists within speech, **the extralinguistic information indicating the presence of pathologies** is of great relevance for this thesis' purposes. The relationship of this trait with other extralinguistic factors such as age and sex is to be explored too.

1.3 NORMAL AND PATHOLOGICAL SPEECH

As it has been discussed previously, *speech* is accomplished through complex articulatory movements that mould the vocal excitation source in order to convey spoken sounds. In this process, three components can be identified: The *excitation source* (be it voiced, unvoiced, a mixture of both or its absence -such as in a pause-) providing the driving force for the production of spoken sounds, the *articulation* defined by the movements of the speech articulators giving form to the production of a certain sound, and the *fluency* defining the rate at which the speech is generated.

Several pathologies might alter one or various of these components, resulting in an impaired production of speech. In reference to the excitation source, it is of particular interest the study of the voiced sounds, since dis-

orders affecting morphological structures of the vocal folds are more easily perceived by studying their vibrational patterns. Having mentioned this, a *speech disorder* can finally be defined as an impairment of the articulation of speech sounds, fluency and/or voice [15, 16]. A succinct description of each one of these types of disorders is presented next [16]:

- **Articulation disorders** are characterized by the production of defective speech sounds and sound combinations that may be distorted, omitted, substituted or added as accessory sounds. These problems encompass many kinds of articulatory defects stemming from faulty learning and/or habits of misuse, or problems related to structural deviations. Other articulation alterations arise from neuromotor disorders that affect the intelligibility of the speech. The most prominent examples are *dysarthria* and *apraxia*. The first involves an impairment in the control and execution of speech movements due to muscle weakness, slowness, incoordination, or altered muscle tone, whereas the latter represents an impairment in the programming of speech movements in the absence of the muscle impairments associated to dysarthria.
- **Fluency disorders**, also known as *stuttering* or *stammering*, describes an impairment in the flow of speech. It is characterized by the repetition of sounds, syllables, words, or phrases; sound prolongations; atypical pauses; word substitution; and use of word fillers that characterize dysfluent behaviour.
- **Voice disorders** are characterized by the abnormal production and/or absence of vocal quality, pitch, loudness, resonance, and/or duration, which is inappropriate for an individual's age and/or sex [15].

It has to be remarked that **only voice disorders are relevant for the interests of this thesis**, and thus fluency and articulation disorders are not considered any further. For this reason, a deeper discussion about voice impairments is to be presented in Section 4.

1.4 DISCUSSION

Some concepts related to speech have been presented in this section. Firstly, a discussion about the subsystems implicated in the production of speech has been made. Then, the source-filter model and some mathematical tools that are often utilised in the analysis of speech signals are introduced. Next, some considerations regarding the linguistic and non-linguistic information that is embedded in speech are discussed. Finally, speech disorders are defined in terms of articulation, fluency and voice impairments. It has to be noticed that amongst all the informative content comprised in speech, **this thesis centres its efforts in the study of the extralinguistic component of pathology** and its relationship with other non-linguistic traits. Within this topic, only **disorders affecting voice** are studied.

PROBLEM STATEMENT

By virtue of the valuable information contained in speech (both linguistic and non-linguistic) and the inexpensiveness and easiness of capturing speech signals in a non-invasive manner, a great deal of interest has arisen in designing automatic systems capable of isolating a certain dimension (or trait within a dimension) for detection purposes. As a matter of example, the literature presents some systems that have focused in extracting extralinguistic information to automatically determine the identity of speakers (speaker recognition) [17], age [18] or sex [19]. In the same manner, paralinguistic information has been extracted to identify the speaker's emotions (emotion recognition) [20] or level of interest [21]. Likewise, the linguistic information has been used to recognize the accent, dialect [22] or the message itself (speech recognition) [4]. Nonetheless, the design of the aforesaid detection systems is challenging due to the complexity of isolating the trait of concern from other overlapping non-target traits. In addition, it is possible that within-trait (intra-trait) or between-trait variations (inter-trait) hinder the performance of detection systems.

Within the field of automatic analysis, one application that has been gaining increasing popularity during the last years is in the **analysis of speakers' condition**. It has been acknowledged that systems which automatically detect, identify and assess pathologies by means of speech present potential advantages compared to traditional detection and evaluation procedures. In particular, they provide an objective assessment of the clinical state of patients, reduce the evaluation time and the cost in diagnosis and treatment [23]. This is important given the large incidence of voice pathologies, and the effects of impairments in the quality of life of patients and in the productivity of the organisations where spoken language plays an important role. In this respect, the lifetime prevalence of voice disorders occurs in almost 50% of population [24]. According to the *Spanish Association of Otolaryngology and Neck and Head surgery*, the Spanish population suffers from at least one voice disorder per year caused by cold and overuse, with a proportion of about 5% to 7% of which require the intervention of an specialist. In the United States, an estimate of 7.5 million individuals suffer from voice disorders caused by overuse, upper respiratory infections, vocal fold lesions, laryngeal cancer and other laryngeal pathologies [25]. Furthermore, there are certain professions that increase the risk of suffering from voice impairments, including singers, factory workers, teachers, clerical workers, salespersons, etc. [26]. Indeed, in the United States 25% of working population

have jobs that require the use of voice, and 3% have occupations in which their voice is necessary for public safety [25]. With this in mind, and due to its relevance, the **automatic detection/assessment of voice pathologies** is the topic on which the current thesis centres its efforts.

The main goal of this thesis is to design automatic systems that provide information about the vocal quality of patients by means of the analysis of voice or speech signals. The system should provide reliable decisions about whether the patient has a voice pathology, and an assessment about the severity of the disease according to a perceptual evaluation scale. Likewise, it should neglect the influence of other overlapping traits/dimensions of speech such as age and sex. Multiple sets of characteristics aiming at describing the phenomena described by pathological voice are to be tested, intending to find the combination of features that provides the best performance in automatic detection and assessment tasks. A discussion about the influence of different variability factors affecting automatic systems, such as sex, age, type of acoustic material, etc. is also performed. Techniques that are typically employed in speaker recognition are tested as well, aiming to compensate for all the variability factors affecting the automatic classification systems. Moreover, the idea of *hierarchical systems* or tree-like structures on which a certain non-target trait is employed to simplify the pathology detection task is also explored. Finally, a generalist assessment system is also designed, trying to provide ratings that resemble those given by a human evaluator, as well as evaluations that account for the continuous nature of the assessment task. Having this in mind, the following sections discuss aspects referent to pathology detection and assessment, trying to respond to the following questions:

- How are pathology detection/assessment systems affected by variability factors?
- What speech production tasks and which characteristics are the most relevant for the automatic detection/assessment?
- How should speech be analysed to capture the peculiarities of pathology while disregarding other non-informative traits found in speech?
- Is it possible to compensate for the variability introduced in speech and affecting the performance of automatic classification systems?
- Can hierarchical classification simplify the detection of pathologies?
- Can the assessment capabilities of a human evaluator be modelled through signal processing techniques?
- Is it possible to account for the continuous nature of perceptual evaluation scales through regression techniques?

OBJECTIVES

HYPOTHESIS:

Voice and speech recordings contain information about the physical condition of all the subsystems in charge of the production of speech. By means of signal processing techniques it is possible to extract relevant information about the prevalence of voice pathologies, as well as its degree of affectation. Thus, it is possible to design a system that automatically analyses voice quality and provides decisions about the condition of patients.

GENERAL OBJECTIVE:

- To design an automatic voice pathology detection and assessment system using voice and/or speech recordings, that takes into account the intrinsic variability of the speech and characterizes the mechanisms of voice pathology, using methods based on speech technologies.

SPECIFIC OBJECTIVES:

- To analyse the factors affecting the performance on automatic detection/assessment of voice pathologies and which are the result of the intrinsic variability of the speech.
- To study the best set of features capable of describing voice pathology while neglecting other non-informative traits contained in speech.
- To investigate the performance of classifiers that compensate for variability factors -such as those typically used in speaker recognition and based on models of mixtures- in an automatic voice pathology detection system.
- To model the perceptual capabilities of a human evaluator through the employment of speech technologies.
- To study the impact of hierarchical systems through the analysis of other extralinguistic cues which might be employed for simplifying the pathology detection/assessment task.
- To design assessment systems capable of providing reliable decisions about voice quality, while accounting for the non-continuous nature of the evaluation task.

Part II

VOICE PATHOLOGY & AUTOMATIC VOICE QUALITY ANALYSIS

VOICE DISORDERS

To address the concept of *voice disorder*, it is first necessary to discuss the properties of a "normal" voice. This, however, poses numerous difficulties since there exist several types of "normality", and the distinction from what can be considered healthy or abnormal relies on subjective perceptual judgments made by listeners or by the speaker itself [27]. Indeed, a singer who uses a deviant voice as a trademark, might acknowledge his/her voice as normal, but this can be perceived otherwise by some listeners. In spite of that, there are certain common characteristics that can be regarded as normative, and thus, are to be utilised as synonyms of *non-pathological vocal condition* in this thesis. Following the definition presented in [25], a normal voice -from now on **normophonic voice**- is characterized by the following properties: (i) a pleasant quality, with an absence of noise, inappropriate breaks, perturbations or atonality; (ii) a pitch in accordance to the age and sex of the speaker; (iii) loudness that is appropriate to the communication event, not so weak as to be unintelligible, not so loud that it calls the attention to itself; (iv) pitch and loudness variations that are available to express emphasis, meaning or subtleties indicating individual feelings and semantic differences; (v) sustainability to meet social and occupational needs. By contrast, **dysphonic voices** are described by an abnormal production and/or absence of vocal quality (including resonance), pitch, loudness, and/or variability which is inappropriate for an individual's age and/or sex [15, 25].

Voice impairments might arise due to misuse, infections, physiological or psychogenic causes, or due to the presence of other systematic disorders (including neurological), vocal abuse, surgery, trauma, congenital anomalies, irradiation, chemicals affecting vocal folds, etc. [16]. Phenomenologically, voice pathologies are manifested in the form of irregular vibration patterns of the vocal folds that might include paralysis, a larger or faster excursion of one of the vocal folds, defects in glottal closure, dissimilar vibratory patterns along one or both folds, variations in vibratory periods and amplitudes of consecutive glottal openings, etc. [16].

The categorization of voice disorders is a challenging task, with literature reporting diverse -and sometimes contradictory- typologies. For instance, according to the aetiology (the hypothesis explaining the origin of the disorder), some authors describe impairments of organic, functional or multi-aetiological type [25]. Other authors employ the terms organic, functional, and neurological, to encompass factors which differ from the previous categorization [28]. This lack of consensus is troublesome, hindering the repro-

ducibility and consistency among different investigations. Bearing this in mind, and for the sake of clarity, the categorization preferred in this thesis is presented in the following subsection.

4.1 CATEGORIZATION OF VOICE DISORDERS

According to their **aetiology**, voice disorders are classified into the following categories [25]:

1. *Organic*: caused by structural or physiological disease, either a disease of the larynx itself or by remote systemic or neurological diseases that alter laryngeal structure or function.
2. *Functional*: caused by psychoneuroses, personality disorders or faulty habits of voice use, in spite of an unaffected laryngeal anatomy and physiology. Also known as *psychogenic* or *non-organic*
3. *Multiple aetiology*: vocal disorders that encompass more than one aetiological group. For instance a psychogenic disorder due to voice misuse might result in an organic disease like vocal polyps.

Another categorization focusing on **kinesiological** characteristics (regarding the vocal fold movements) divides voice disorders according to whether the vocal folds over-adduct (the glottis closes too tightly) or under-adduct (incomplete glottal closure), generating the following categories [25]:

1. *Hypofunctional*: vocal folds are under-functioning and have inadequate tension, so air escapes through.
2. *Hyperfunctional*: vocal folds are overly tense and compress too tightly together.

Finally, other definition focusing on **perceptual** characteristics divides disorders into [16, 29]:

1. *Dysphonia*: abnormal voice production according to the perceptual parameters of *pitch*, *loudness*, *quality* and *variability*.
2. *Muteness*: inability to produce sounds or articulating speech.

This perceptual categorization is of great importance in pathology detection tasks due to its direct relationship with acoustic indicators extracted from the voice, and therefore is of great significance for the objectives of this thesis. Considering that, the following section presents a deeper discussion of voice impairments in terms of the perceptual categorization. Additionally, further subcategories are introduced according to the aetiology following the pathology. It is important to remark that muteness is omitted from the discussion, as it is related to the absence of voice or speech and thus cannot be analysed by means of vocal recordings.

4.2 PERCEPTUAL CATEGORIZATION OF VOICE DISORDERS

Voice disorders might affect the perceptual parameters of pitch, loudness, quality and variability. The following paragraphs describe the disorders in terms of these perceptual dimensions. In addition, and when possible, some illustrations of vocal folds pathologies are presented. In all cases, the pictures are referred to organic impairments since functional disorders do not affect larynx structures. It is also worth mentioning that some pathologies are multidimensional, and hence might alter one or more perceptual parameters. For instance, a pathology such as the Reinke's oedema influences simultaneously the pitch, loudness and quality perceptions.

4.2.1 Disorders affecting pitch

Pitch is the perceptual correlate of f_0 , and is mainly determined by the mass and elasticity of the vocal folds in relation to their length. Pitch disorders are described by an inconsistently higher or lower sounding voice according to the individual's context, or when the sound is tremulous, bizarre or monotonous. According to the aetiology, pitch impairments are further divided as [16]:

1. **Functional:** when the voice does not correlate with what is expected of the speaker's physiology. For instance, in *mutational falsetto* an abnormally high-pitched voice is produced by an adult male (not correlating with his age and sex). Likewise, in *ventricular phonation* the speaker produces an abnormally low-pitched voice through the vibration of the false vocal folds.
2. **Organic:** in some conditions such as *laryngeal web* (a membrane-like malformation that extends across the larynx), *underdeveloped larynx* or *structural asymmetry* (abnormal approximation of the vocal folds) produce an unexpectedly high-pitched voice. By contrast, an atypical low-pitched voice is particular of disorders such as *Reinke's oedema* (accumulation of fluid or swelling, usually in both vocal folds), *virilization* (masculinization of the female voice after hormone therapy), *glottalization* (low-frequency popping or ticking sound) and *tremulousness* (pitch deviation). An example of a larynx affected by a disorder affecting pitch (Reinke's oedema) is presented in Figure 4.1.

4.2.2 Disorders affecting loudness

Loudness is the perceptual correlate of the intensity of the sound pressure created by the release of air through the glottis. Disorders of loudness arise when the voice is louder or softer than expected in relation to the place and circumstances on which it is produced, or when there exist loudness variations that are not appropriate for communication. Loudness disorders are divided according to their aetiology into [16]:

Figure 4.1. Reinke's oedema, a pathology characterized by a swelling of the vocal folds.

1. **Functional:** two main causes are considered. On one hand, voices that are too loud or too soft are often an indicator of personality disorders, e.g., overly aggressive, shy, or socially insecure behaviour. On the other, because of environmental stress, people are required to speak loudly in their occupations, often creating laryngeal trauma with subsequent changes in the vocal organs and consequent voice disorders.
2. **Organic:** within this category diverse pathologies are described. For instance, *bowed vocal folds* are muscular deformations stemming from long-term heavy use of the voice, particularly in elderly patients. Other disorder is *paresis* (or vocal fold paralysis) which is referred to slowness, weakness, or absence of complete glottal closure due to paralysis of one (unilateral) or both (bilateral) vocal folds. One example of a larynx affected by paresis is illustrated in Figure 4.2.

Figure 4.2. Example of paresis and its comparison with the normal function of the vocal folds during respiration and phonation. Image retrieved from [30]

4.2.3 Disorders affecting quality

Quality is the perceptual correlate of complexity and is described by the pattern of vibration and resonance of the vocal folds. From the perceptual point of view, there are four elements to consider in a pathological voice in terms of quality [16, 25, 29]:

1. **Strain:** corresponds to an excessive tension in the larynx which may result in over adduction of vocal folds, creating disturbances in its vibratory patterns. This is usually correlated to vocal problems of functional origin, but organic problems such as oedematous vocal folds, and any other structural alteration may increase strain perception.
2. **Breathiness:** related to turbulent air streams released during incomplete vocal closure. Several organic conditions modify the glottal configuration to create an incomplete glottal closure associated with a breathy voice; for instance, *muscular tension dysphonia* (simultaneous contraction of laryngeal adductor and abductors), *paresis*, growths protruding into the glottal space (*nodules*, *polyps*, etc.), or disorders that affect glottal resistance. An example of impairments affecting the perception of breathiness are presented in Figure 4.3.
3. **Harshness or roughness:** phenomenon expressing irregularities or vibration defects caused by alterations of the ordinary, repetitive, synchronous vibration of the vocal folds. Random aperiodicity may include any disease or condition in the larynx that changes the size, stiffness, or surface characteristics of one or both vocal folds (laryngitis, nodules, polyps, etc.) or that causes excessive squeezing of the folds (e.g., pressed phonation by a patient with contact ulcers), etc. Pathologies affecting the roughness perception are illustrated in Figure 4.3.
4. **Resonance:** the two most common resonance defects are due to too much nasal resonance (hypernasality) or insufficient nasal resonance (hyponasality). The first is caused by incomplete closure of the velopharyngeal valve or by a fistula in the structures separating the oral and nasal spaces, and the second is caused by blockage of the nasal passageway. One pathology affecting the resonant patterns of the vocal tract is the cleft lip and palate, which is illustrated in Figure 4.4.

Figure 4.3. Disorders affecting a complete closure of the vocal folds: (a) nodules and (b) polyps.

4.2.4 Disorders affecting variability

Variability is the perceptual correlate of the variations in pitch, loudness and quality parameters. Normophonic voice possesses adequate variability

Figure 4.4. Cleft lip: pathology affecting the resonant patterns of the vocal tract.
Image retrieved from [31]

during contextual speech to convey subtle, intellectual and emotional meanings. In voice disorders these fluctuations may be flattened, excessive or inconsistent.

4.3 DISCUSSION

This section has introduced some of the terminology to be used during the development of this thesis, including the concepts of normophonia and dysphonia. Due to the sometimes contradictory categorisations that are often presented in literature, this section has introduced the taxonomy that is preferred in this document. A great deal of interest has been put on the perceptual categorisation of disorders regarding the correlates of pitch, loudness, quality and variability. In addition, and with respect to the alterations of vocal quality, a further discussion has been provided to address the characteristics of strain, breathiness, roughness and resonance.

AUTOMATIC VOICE QUALITY ANALYSIS SYSTEMS

The clinical evaluation of voice often relies on an instrumental examination and a perceptual analysis of the speech. In this respect, the *instrumental* medical examination focuses on a primary aetiological diagnosis through the investigation of acoustic, aerodynamic, electroglottographic, videolaryngostroboscopic and/or the exploration of other types of biosignals; whereas the *perceptual* examination extracts multidimensional information, that is not quantifiable instrumentally, by means of a qualitative description of the perceived degree of dysphonia that is present in the voice [32–34]. This information might be complemented by an interview where the patient states his/her symptoms, the examination of his/her medical records, evaluation of other body functions and systems, and exploration of the laryngeal structures and their function. These procedures should lead the speech pathologist to a diagnosis about the condition of the patient. The diagnostic process is differential, i. e., all possible causes of a problem are considered, and then the available information is matched against each one of the hypothesis explaining the disorder in the search for a match [35]. In some severe pathological cases a decision about the condition of the patient is straightforward, but in others, it would probably be conditioned by subjective factors or the observations and hypothesis made by the clinician. The increasing need of improving the diagnosis of voice pathologies has given rise to an emerging field of analysis called *Automatic Voice Quality Analysis* (AVQA), that aims at analysing, classifying and quantifying the degree to which a patient is affected by a voice disorder. This analysis is performed using automatic systems that provide objective measurements of the patient's vocal condition, exploiting the close relationship that exists between acoustic features extracted from the speech and voice pathology [36]. This reduces the evaluation time and the cost of diagnosis and treatment, providing extra advantages such as the avoidance of invasive procedures thanks to the employment of speech signals which are easily recorded by inexpensive means. Generally, AVQA systems follow a common structure on which characteristics are extracted from the acoustic signal in the form of a set of parameters to accomplish a further decision making task. A typical AVQA scheme is presented in Figure 5.1.

It is important to remark, though, that two initial considerations -referred to the *input speech* and *decision* blocks in the AVQA depiction- should be addressed to respond to the following questions: (i) what type of speech task is to be used for the design of the system; (ii) what type of decision should

Figure 5.1. Depiction of a typical AVQA system.

the system provide. The current chapter is devoted to discuss these two aspects of the AVQA design; whereas in the theoretical background (Part iii), a detailed explanation about of each one of the constituting blocks of Figure 5.1 is provided.

5.1 SPEECH PRODUCTION TASKS

The interest of considering the acoustic material in the design of AVQA systems arises from the fact that depending on the type of utterance, different configurations of the speech production subsystems (articulatory, phonatory, respiratory, etc.) stem. In addition, some pathologies are more likely to be identified when examining determined units of speech. In general, two types of *speech production tasks* are examined for the evaluation of vocal quality due to their relevance in dysphonic voice analysis. These tasks are:

- *Sustained phonation*: the result of the phonation of voiced sounds due to the vocal folds vibration, as when a vowel is uttered and maintained during a certain amount of time. Phonation exhibits a large range of behaviours, ranging from periodic or regular vibration, to aperiodic (irregular) vibration or sounds with no apparent vibration at all [37].
- *Running or connected speech*: the result of the source signal (either voiced, unvoiced, a mixture of both or its absence) being modulated by the articulatory subsystems, as when speech is being produced while uttering a certain word or pronouncing sentences.

Using one type of speech production task or the other poses certain advantages or disadvantages that should be considered beforehand. The following subsections describes both type of tasks more deeply.

5.1.1 Sustained phonation:

The sustained phonation of vowels constitutes the most commonly employed material in AVQA systems. Some of its advantages include [38, 39]:

- The facility to be analysed by automatic tools.
- Vowels are easily produced by speakers.
- Vowels are good descriptors of pathology as most of the laryngeal disorders mostly affect vocal folds, and hence alter normal phonation.

- Sustained phonation is not affected by paralinguistic or extralinguistic characteristics such as speaking rate, speaker's dialect, intonation, and idiosyncratic articulatory behaviour.
- Vowels often generate simpler acoustic structures that might lead to consistent and reliable perceptual judgements of voice quality.
- Vowels do not depend of extra processing stages (such as voiced/unvoiced detectors) for the design of AVQA systems.

The selection of the vowel to be uttered is also a relevant matter. For instance, certain open vowels, such as /a/, are produced with a relatively open tract allowing the examination of the entire vocal tract apparatus. By contrast vowels such as /i/ and /u/, may not allow this examination due to the striking separation between the front and back cavities of the mouth which is made in their production [40]. Besides that, the type of vowel -along with the vocal effort and the muscle tension in the larynx- influences the degree of vocal folds approximation, and affects the voice quality perception [41]; and as such, should be considered when designing AVQA systems.

5.1.2 Running Speech

Despite the running speech task is not as widely popular in AVQA systems as the one based on sustained vowels, there exist strong arguments favouring its use. One interesting property of running speech comes from a phenomenon called *coarticulation*, which is related to the influence of the preceding and succeeding speech sound on the unit that is currently uttered. The dynamical effects introduced by coarticulation might be relevant for certain applications. Besides that, it has been stated that the impression of certain characteristics of vocal quality are more easily perceived on vowels generated in a voiced context, vowels after a glottal closure, or during the production of strained vowels [24]. Some additional advantages reported in literature in favour of analysing connected speech are the following [36, 42]:

- Running speech requires switching on and off the vibration of the vocal folds continually, or that the voicing is maintained while the supraglottal apparatus changes. This might facilitate the exploration of certain dynamic aspects in the speech production.
- Speakers are less likely to compensate for their voice problems while producing connected speech than while phonating sustaining sounds.
- Running speech provides a more realistic scenario since sustained phonation is more characteristic of singing rather than speaking.
- Running speech contains fluctuations of vocal characteristics in relation to voice onsets, terminations and breaks, which are considered crucial in voice quality evaluation. These characteristics are not fully represented in short registers of phonation such as in a sustained vowel.
- In certain voice disorders (e.g. spasmodic dysphonia), the production of sustained vowels is less symptomatic than in connected speech, which may lead to an underestimation of the impairment.

- Running speech contains variations in pitch and loudness, parameters that are important in the analysis of abnormal voice quality.

5.2 DECISION TASKS

As it has been stated previously, the type of decision making task influences the design of the AVQA systems. In this regard, three fundamental tasks may be considered: voice pathology detection, voice pathology identification and voice pathology assessment.

5.2.1 *Voice pathology detection and identification:*

Voice pathology detection is a two-classes decision making process aiming to decide whether a given speech register is normophonic (if it is normal in terms of quality), or pathological (if it suffers from vocal pathology or dysphonia). On the other hand, *voice pathology identification* is a multi-class decision making process on which the goal is to assign a category to the input speech. The identification task is typically made in terms of the actual pathology (nodules, Reinke's oedema, etc.), the aetiology (organic, functional, etc.) or any other categorisation that groups general aspects of the analysed speech. From a practical point of view, identification is more challenging than detection, because of the multiple-class scenario on which it is defined. Nonetheless, both tasks are intricately complex due to several factors, such as the wide range of profiles that are found for normophonic voices, the documented overlap between normophonic and pathological voices [43], the close relationship between disease and certain quality factors due to normal processes such as ageing [44], the simultaneous presence of pathologies of different aetiology in the same patient, etc.

5.2.2 *Voice pathology assessment*

Voice pathology assessment is aimed at grading the level of pathology that is perceived in a given speech signal. This is of great relevance since, as stated before, it is not possible to instrumentally delimit a phonation behaviour categorically. An useful descriptor of dysphonia is the **hoarseness**, which portrays the noisy, atonal and/or odd vocal resonance patterns encountered in voices [25]. The hoarseness is widely employed in literature, as perceptually, it is described as a superclass that contains roughness and breathiness -the two most reliable traits describing vocal quality- [24]. The assessment task is generally performed in concordance to a perceptual rating scale that evaluates voice quality and provides information about the level of dysphonia. While the perceptual scales have been designed to evaluate every aspect that is relevant to voice quality, the reliability of the ratings is conditioned by the multidimensional aspects of voice quality, the intrinsic variability of speech, the subjectivity of perception [45], and the nonlinear relationship between pathology and measured or perceived voice quality [43]. Such aspects

should be taken into account when designing automatic voice pathology assessment systems.

Perceptual ratings are often made using scales that describe the severity of the speech disorder. Some popular perceptual evaluation scales include the *Grade-Roughness-Breathiness-Asthenia-Strain* (GRBAS), *Voice Handicap Index* (VHI) and *Consensus Auditory-Perceptual Evaluation of Voice* (CAPE-V). Another popular evaluation tool is the so called *hoarseness diagram*.

5.2.2.1 GRBAS perceptual scale

GRBAS is a simple and reliable scale for evaluating voice quality. It is composed of five descriptors ranging from 0 to 3, where 0 refers to a healthy condition, 1 to light disease, 2 to moderate impairment and 3 to grave disorder. The five descriptors are detailed next [33, 46]:

1. *Grade* ($\mathcal{G}$), a measure of global voice quality or hoarseness.
2. *Roughness* ($\mathcal{R}$), linked to the level of perceptually audible impressions of irregular glottal pulses, abnormal fluctuations of f_0 or multiple impulses perceived separately (dyplophonia or voice breaks).
3. *Breathiness* ($\mathcal{B}$), related to the perceived loss of air due to an incomplete glottal closure and which might include aphonic moments.
4. *Asthenia* ($\mathcal{A}$), associated to weakness or lack of power in voice.
5. *Strain* ($\mathcal{S}$), linked to the perceived level of vocal effort.

A simplified scale limited to $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ is frequently found in the literature due to the unreliability of the $\mathcal{A}$ and $\mathcal{S}$ parameters [24]. This reduced scale resembles the *Roughness-Breathiness-Hoarseness* (RBH) rating scale used in German-speaking countries [47], and encompasses the most important information for voice quality assessment.

5.2.2.2 Consensus Auditory-Perceptual Evaluation of Voice

CAPE-V is not only regarded as a rating scale resembling GRBAS, but it is also a protocol itself [48, 49]. In this respect, CAPE-V recommends a procedure which includes the recording of the phonation of the sustained vowels /a/ and /i/ three times; six predefined sentences with different phonetic contexts; and a conversation in response to the standard question "Tell me about your voice problem" [48]. Then, voice quality is described by means of 6 traits: *overall severity*, *roughness*, *breathiness*, *strain*, *pitch* and *loudness*, which are rated using a 100-millimetre line resembling a visual analogue scale, to avoid the use of discrete points [48].

5.2.2.3 Voice Handicap Index

VHI is proposed as a self-assessment tool for measuring the psycho-social handicapping effects of voice disorders [50]. It consists of 30 items, from which 10 correspond to a physical, 10 to an emotional, and 10 to a functional dimension. The *physical dimension* represents the self-perception of laryngeal

discomfort and voice output characteristics; the *emotional dimension* indicates the patient's affective responses to the voice disorder; whereas the *functional dimension* includes statements describing the impact of the voice impairment in their daily activities. Each question has 4 possible answers ranging from "never" (scored as 0) to "always" (rated as 4). After a patient has been intervened, 18 points ensure -with a 95% confidence- that changes in total scores are due to the treatment and not to the **VHI** variability. A graphic indicating some of the items in the **VHI** scale are presented in Figure 5.2.

Appendix

Voice Handicap Index (VHI), Henry Ford Hospital

Instructions: These are statements that many people have used to describe their voices and the effects of their voices on their lives. Circle the response that indicates how frequently you have the same experience.

- F1. My voice makes it difficult for people to hear me.
- P2. I run out of air when I talk.
- F3. People have difficulty understanding me in a noisy room.
- P4. The sound of my voice varies throughout the day.
- F5. My family has difficulty hearing me when I call them throughout the house.
- F6. I use the phone less often than I would like.
- E7. I'm tense when talking with others because of my voice.
- F8. I tend to avoid groups of people because of my voice.
- E9. People seem irritated with my voice.
- P10. People ask, "What's wrong with your voice?"
- F11. I speak with friends, neighbors, or relatives less often because of my voice.
- F12. People ask me to repeat myself when speaking face-to-face.
- P13. My voice sounds creaky and dry.
- P14. I feel as though I have to strain to produce voice.

- E15. I find other people don't understand my voice problem.
- F16. My voice difficulties restrict my personal and social life.
- P17. The clarity of my voice is unpredictable.
- P18. I try to change my voice to sound different.
- F19. I feel left out of conversations because of my voice.
- P20. I use a great deal of effort to speak.
- P21. My voice is worse in the evening.
- F22. My voice problem causes me to lose income.
- E23. My voice problem upsets me.
- E24. I am less outgoing because of my voice problem.
- E25. My voice makes me feel handicapped.
- P26. My voice "gives out" on me in the middle of speaking.
- E27. I feel annoyed when people ask me to repeat.
- E28. I feel embarrassed when people ask me to repeat.
- E29. My voice makes me feel incompetent.
- E30. I'm ashamed of my voice problem.

Note. The letter preceding each item number corresponds to the subscale (E = emotional subscale, F = functional subscale, P = physical subscale).

Figure 5.2. Example of the **VHI** test. Graphic taken from [50].

5.2.2.4 The hoarseness diagram

The hoarseness diagram is a two-dimensional representation describing the noise content of voice registers. The hoarseness diagram uses abscissa to represent irregularity (perceptually correlated to roughness); whereas the ordinate represents the additive noise content of voices (correlated to breathiness) [51, 52]. An example of the hoarseness diagram for a normophonic and a pathological voice is illustrated in Figure 5.3.

Figure 5.3. Hoarseness diagram for a normophonic and a pathological voice.

5.3 DISCUSSION

The current section has introduced some concepts related to [AVQA](#) systems. A typical [AVQA](#) scheme has also been presented and two aspects of its design have been discussed: the type of acoustic material, and the expected decision outputted by the system. In this regard, the acoustic material resulting from the sustained phonation and running speech production tasks has been described and discussed in terms of the advantages and disadvantages of using one or the other. Finally, the decision making tasks outputted by the [AVQA](#) methodology are also introduced. In particular, three decision making tasks are defined: voice pathology detection, identification and assessment. With respect to the latter, a list of some common assessment rating scales has also been introduced. This thesis considers **the influence of using both types of speech tasks in [AVQA](#) systems**. However, it is limited to designing **voice pathology detection** and **voice pathology assessment** systems only.

VARIABILITY ASPECTS AFFECTING VOICE QUALITY ANALYSIS SYSTEMS

The variability embedded in speech has long been recognised as a major source of errors in automatic classification systems based on speech. For instance, several variability factors identified in the design of speaker recognition systems are described in [53]. Translated to terms of AVQA systems, these variability factors are the following:

- *Peculiar intra-class variability*: manner of speaking, age, sex, inter-session variability, dialectal variations, emotional condition, etc.
- *Forced intra-class variability*: Lombard effect, external-influenced stress, cocktail-party effect, etc.
- *Channel-dependent external influences*: type of microphone, bandwidth and dynamic range reduction, electrical and acoustical noise, reverberation, distortion, etc.

It is worth noting that the forced intra-class variability is more common of unsupervised recording environments rather than from controlled clinical settings. That does not imply, though, that their effects should be disregarded. For instance, they are of significant importance in telemedicine scenarios where the recording conditions might vary widely. Despite that, and for the purpose of simplicity, the forced intra-class variability is to be omitted from further discussions, and the term *intra-class variability* is to be referred to the *peculiar intra-class variability* only. Moreover, the intra-class and channel-dependent factors can be further associated to the dimensions of speech introduced in Section 1.2. In this manner, the intra-class variability is related to the linguistic, paralinguistic and extralinguistic spheres, while the channel effects encompass the transmittal dimension.

Having this in mind, the current chapter introduces some factors that are of interest in the design of AVQA systems. Because some of these aspects cannot be readily controlled, some final comments are presented at the end of this chapter regarding which variability factors are to be investigated.

6.1 INTRA-CLASS VARIABILITY

One important aspect to be outlined in any speech-related system is the effect of the intra-class or *intersession* variability. In a speaker recognition

system, aiming at recognising the identity of a target speaker, the intersession effect might be described by the differences arising between recordings of the target speaker due to vocal effort, physical or emotional condition, etc. In the AVQA field, the intersession variability might be explained by the acoustic diversity among different pathologies, the sex or age of the speaker, or the spurious information introduced by other linguistic, paralinguistic or extra-linguistic effects. In this regard, several intra-class factors that might affect the performance of voice pathology classification systems include *linguistic* aspects such as the speech production task (already discussed in Section 5.1), the dialect and accent of the speaker; *paralinguistic* events such as the emotion or the vocal effort; or *extralinguistic* effects such as the sex or age of the speaker. Bearing this in mind, this section describes some of these intersession variability factors affecting AVQA systems.

6.1.1 *Dialects and accents*

Dialects are the result of systematic, internal linguistic changes that occur within a language, reflected in the form of structural alterations in phonology, morphology, syntax, lexicon or semantic [44]. Dialects have been identified as an important aspect when defining communication disorders. Undoubtedly, not accounting for dialect features may result in the misdiagnosis of communication disorders [44]. For instance, several key features of African-American English phonology have been found to overlap with identifiers of speech sound delay/disorder in the phonology of general American English, making the distinction from normal and disordered states problematic in African-American speakers [54]. This phenomenon has been found in other contexts where non-prestige social dialects are often incorrectly associated to disordered speech [44]. *Accents*, on the other hand, are linguistic changes within a language that occur mainly at the phonological level. It has been long identified as an important confounding source in speech-related applications. For instance, accent is described as the most important source of variability between speakers in speech recognition systems in [55]. A further study presented in [56], confirmed that accent degrades classification rates, with errors increasing around 40-50% in cross-accent speech recognition scenarios. In general, it has been found that performance degrades when recognizing accented from non-native speech [57].

Above facts suggest that accent and dialects should be accounted when designing AVQA systems using running speech, as these might degrade performance.

6.1.2 *Vocal effort and loudness*

There exist evidences indicating that loudness and vocal effort alter perceptual and acoustic parameters extracted from speech, and therefore might impact AVQA systems. Phenomenologically, variations in vocal effort affect the shape of the glottal pulses, changing the closing velocity waveform and

affecting the relative duration of the closed interval [58]. In voices produced with increased vocal effort there have also been found significantly greater values in parameters such as subglottal pressure, translaryngeal airflow, and maximum flow declination [59]. Similarly, it has been stated that the medial compression of the vocal folds is enhanced when vocal effort is augmented, which results in an improved glottal closure, enlarged vocal intensity, and increased f_0 and amplitude [60]. Vocal effort also alters the duration of vowels, consonants, and the pausing behaviour during speech production [58]. In terms of quality, voice produced at excessive vocal effort (loudness) is perceptually described as creaky [58] or strained [59]. Not surprisingly, this has consequences on parameters extracted from speech. For instance, jitter, shimmer, *Normalized Noise Energy* (NNE) and two electroglottographic parameters have been found to vary significantly among vowels produced at three vocal effort levels (low, normal, high) [61]. Similarly, it has been reported that jitter and shimmer significantly increased their value with decreasing voice loudness [62], being also identified as one of the most important factors influencing the computation of these perturbation parameters, alongside with the sex of the speaker and the type of uttered vowel [63]. Other sets of parameters which are impacted include cepstral features, which have been reported to differ substantially at different loudness levels [60]. This has also been confirmed in [59], where significant differences arose in 4 aerodynamic and 2 cepstral measures when comparing phonation at different effort levels. Authors in [64] have investigated the effects of increased loudness in pathological phonation. Results indicate that louder voicing reduces the values of perturbation parameters in normophonic speakers or in superficial vocal fold pathologies, while in cancer or vocal fold paralysis, louder phonation significantly enhances the irregularity of vocal folds vibration.

The incidence of loudness and vocal effort in the computation of features suggest that its effect might affect the performance of AVQA systems, and should certainly be taken into account in the design process.

6.1.3 Emotion

The study of the emotional content embodied on speech has garnered a lot of attention within the speech research community. Indeed, the term *affective computing* has been coined to describe the automatic sensing, recognition and synthesis of human emotions from any biological modality such as speech or facial expressions [65]. Some of the reasons for this great deal of interest are on the -expected- improvement in human/machine interactions by the incorporation of emotional synthesis and recognition systems. Emotions are typically represented as the combination of the *primary (or archetypal) emotions* of anger, disgust, fear, joy, sadness and surprise [66]. Any emotion can also be portrayed by two subjective descriptors: its *activation*, referred to the amount of energy expressing the emotion; and its *valence*, describing the level of pleasantness or unpleasantness. The value of these two descriptors conditions the emotion that is being elicited. For instance, a high activation (due to the arousal of the sympathetic system) is found in joy, anger,

and fear, and is reflected in the form of loud and fast speech with an augmented energy in the high-frequencies, and an increased average pitch with a wider range. On contrary, low activation emotions (due to the arousal of the parasympathetic system) such as sadness, produce speech that is slow, low-pitched, and with a low energy content in the high-frequencies [66].

There exist some studies considering the effect of emotions in speaker recognition systems. For instance, emotions are regarded as a factor affecting automatic recognition of children's speech in [67]. In [68], the effect of an emotions recogniser previous to a speech classification process is investigated. Authors report that affective speech downgrades the system performance, and that a cascading scheme is highly effective in improving recognition rates. Despite these facts, little is known about the influence of emotion in AVQA schemes, but according to the evidence found in the field of speaker recognition, it might be hypothesized that affective speech is a confounding factor that should be taken into consideration.

6.1.4 Sex

The variability introduced by the sex of the speaker remains as a major concern in the design of speech-based systems. Indeed, authors in [55] reported that this factor accounted for the second most prominent source of variability -after accent- in speech recognition systems. Certainly, literature states that the performance of speech recognition, identification or verification systems improves by employing a-priori information about the sex of the speaker [69]. For instance, authors in [70] obtained a 2% of accuracy improvement in a speaker recognition system when using sex-specific models.

The nature of the variability introduced by the sex of the speaker stands on physiological, acoustic, and psychophysical factors [19]. Regarding *physiological* differences, the human laryngeal anatomy differs between sexes at a variety of levels. Particularly, males tend to have a more acute thyroid angle; thicker vocal folds; a longer vocal tract; a larger pharyngeal-oral apparatus, thyroid lamina and skull compared to that of females [3, 8]. Studies of excised human larynges have shown that anteroposterior dimensions of the glottis are 1.5 times larger in men than in women [71]. Besides that, the female pharynx has been found to be shorter than of males during the production of the three cardinal vowels. This may be a key factor in distinguishing between male and female voice qualities during speech production [8]. In addition, the observation of the glottis during phonation has suggested the presence of a posterior glottal opening that persists throughout a vibratory cycle and which is common for female speakers, but occurs much less frequently among male speakers [72]. Indeed, about 80% of females and 20% of males have a visible posterior glottal aperture during the closed portion of a vocal period [73]. Regarding *perceptual* differences, parameters such as effort, pitch, stress, nasality, melodic patterns of intonation and coarticulation are used for characterising female voices, while male voices are judged on the basis of effort, pitch and hoarseness [74]. It is also argued that female voices possess a "breathier" quality than male voices [73]. The

pitch is the most known trait differentiating sexes [74], with females' pitch higher than of males' by as much as an octave [75]. This pitch difference might influence the perception of dysphonic voices since lower pitch is perceived as rougher [24]. In addition to the pitch, literature reports significant differences between male and female speakers' formants (f_1 , f_2 , f_3 , f_4) [19]. This is because the vocal tract length for males is longer than that of females, producing on average formant patterns scaled upward in frequency by about 20% [74]. There are also several important acoustic consequences of the posterior glottal opening during the closed phase of phonation, which is more frequent in females. A first consequence is a breathier voice quality which is the result of a larger amount of air passing through the vocal tract [3] and that affects the relative amplitude of the first harmonic of the speech spectrum [71, 76]. A second consequence is the widening of the f_1 bandwidth, which is the result of the glottal aperture that produces energy losses particularly at low frequencies [72, 76]. A third acoustic consequence is the generation of turbulence in the vicinity of the glottis [76], perceived with a high level of aspiration noise in the spectral regions corresponding to f_3 , and contributing to a breathier voice quality [75]. A final consequence is a lower spectral tilt due to the presence of aspiration noise [75], which turns out to be a significant parameter for differentiating between male and female speech samples [72].

In addition to the acoustic differences reported from the study of the raw speech, there are some differences in the glottal components among sexes. On one hand, the female glottal waveform tends to have a shorter period, lower peaks and peak-to-peak flow amplitudes than that of males [77]. Likewise, the derivative of the glottal waveform does not present an abrupt discontinuity during the closing time due to the incomplete closure of the vocal folds [19]. In general, it is stated that female glottal components are symmetric, with opening and closing portions of the waveform tending towards equal duration [78]. Conversely, and regarding the glottal waveform of male speakers, it is found that the open quotient is smaller and the maximum flow declination rate is greater than of females [72]. Moreover, the closing portion of the waveform generally occupies 20 – 40% of the total period and it might not exist an easily identifiable closed period [74]. In general, it is stated that male glottal waveforms are asymmetrical and present a hump in the opening phase. Finally, it is worth noting that sex differences are found to be age and hormone-dependent, and thus is of great importance considering the effect of age when studying male or female voices.

6.1.5 Age

According to the male-female coalescence model of ageing voice, hormone-related factors cause changes in voice production systems. In this manner, hormones during puberty are responsible for the differences between males and females in adolescence, but these changes are counteracted to some degree by hormone related factors associated with menopause and ageing [8]. During males' puberty, the thyroid cartilage develops the Adam's ap-

ple, the muscular and mucosal layers of the vocal folds thicken, the vocal folds lengthen and widen, the cricothyroid membrane widens, and the corresponding muscle becomes more powerful [3, 79]. As a result, the fundamental frequency decreases an octave compared to that of a child [79]. During females' puberty, there is little development of the thyroid cartilage or of the cricothyroid membrane, and the vocal muscle thickens slightly but remains supple and narrow. As a result, the female's voice fundamental frequency becomes one third lower than that of a child [79].

The age effects on the larynx tend to be more significant in men than in women. In this manner, males experience an increasing of the fundamental frequency as a result of muscle atrophy, thinning of the lamina propria, general loss of mass and ossification and calcification of the larynx that starts during the third decade of life [8]. In females, ossification and calcification starts in the fourth decade of life, and in some cases never completely ossify. However, due to menopause effects, a lowering of fundamental frequency prior to senescence occurs [8].

6.2 CHANNEL-DEPENDENT EXTERNAL INFLUENCES

These dimensions include all the effects that aggregate variability to speech registers because of the mere act of recording. This is a well-known problem that has long been identified in speech and speaker recognition systems [80]. Several aspects affect the recording process, including the instrumentation (type of microphone, analogue-to-digital converter, etc.), the acoustic environment (office, recording studio, etc.) and the transmission means (land-line, cellular, etc.) [81]. Similarly, background noises, noises made by speakers (such as lip smacks), noise in the input device itself, etc., are recognised as sources that impair performance of speech recognition systems [80]. Another problem that might arise, is the variability introduced because of the mismatch in the recording conditions, between the registers employed for training the models and those used for testing purposes. For instance, as when a certain microphone is used for recording the training utterances, but the model is verified with a different equipment. Indeed, the microphone is expected to modify the speech spectrum, and anything that modifies the spectrum may cause difficulties in recognition tasks [81]. All these external influences do surely affect AVQA systems, and should be taken into account in their design.

6.3 DISCUSSION

This section has presented some variability factors that might affect the performance of AVQA systems. The list is not exhaustive, but is composed of some of the most important variables that -to the author's opinion- ought to be considered in the design of systems that detect/identify/assess voice pathologies. Despite it would be desirable to study the influence of every single variable described in this section, there are some effects that cannot

be controlled within the [AVQA](#) work-flow and thus, have to be disregarded. As a matter of example, emotions are considerably difficult to elicit naturally in laboratory conditions. In fact, actors are often employed in affective speech research. The inability to regulate or to ascertain the emotional state of speakers obligates to discard this trait from further investigation. In the same manner, dialects and accents require a very extensive database containing regional and dialectal differences to correctly model their effects. Unfortunately, this speech material is not available. Indeed, some of the databases employed during this thesis are limited to speakers of certain geographical regions to avoid the introduction of this variability factor. At the end, the variables under control are constrained to the extralinguistic traits of **sex** and **age**, and to the linguistic trait of **type of uttered speech** (running speech or sustained vowels). In the same manner, **channel-dependant effects** should be regarded too, since subtleties in the recording process are expected, including those due to varying distances from mouth to microphone, slight variations in the recording environment or noises introduced during the acquisition process.

Part III

DESIGN OF AUTOMATIC VOICE QUALITY ANALYSIS SYSTEMS

STATE OF THE ART

There exist in literature a wide variety of methodologies related to the detection, identification and assessment of voice pathologies, with results that vary depending on the application, the acoustic material, and the actual methodology that is used. Based on the design blocks depicted in Figure 5.1, the current section discusses each one of its constituting parts in terms of the techniques that are followed in the literature. This includes the databases that are commonly utilised in the design of AVQA systems, some typical preprocessing techniques, dimensionality reduction schemes, as well as the decision machine used for discrimination purposes.

7.1 DATABASES

The acoustic material to be employed in the design of AVQA systems should contain exemplar data describing the classes under study. This is often composed of corpora of speakers portraying the investigated pathologies and normophonic cases. The data acquisition process should follow certain guidelines not to introduce unexpected variability, including the avoidance of external sources of noise or the maintenance of the same acoustic and instrumental conditions during the recording process. Moreover, the corpus should be large enough to contain all possible variations within pathological or normophonic conditions, while trying to maintain a balance among classes in terms of age, sex, dialect, etc. The idea by doing so, is to design generalist models that disregard the particularities related to the speakers themselves, and thus to identify the actual characteristic related to normophonic or dysphonic conditions. The management of the corpus for the storage and accessibility of recordings from a medical perspective should also be taken into account, as this permits the creation of synergies towards certain labours such as the diagnosis of the pathologies, or the assessment from a perceptual point of view as given by different evaluators. Some considerations referred to the management in a clinical setting for a large corpus of dysphonic and dysarthric speakers are discussed in [82].

Literature reports the existence of several public and privative databases that have been recorded for the purposes of detection, identification or assessment of voice pathologies. Regarding *public* databases, the *Massachusetts Ear and Eye Infirmary* (MEEI) corpus [83] is probably the most widely employed, being for years, the sole resource that was available for the study of

pathological speech. *MEEI* contains approximately 700 registers of the vowel /a/ and the first sentence of “*the rainbow passage*” text, recorded at varying sampling frequencies (25 kHz to 50 kHz). A subset of the database -chosen to ensure a balance in age, sex and pathologies- has been published in [84], becoming a standard partition for comparisons among different works in literature. Despite its popularity, *MEEI* suffers from well-known problems which might bias outcomes, like the different recording conditions of normophonic and pathological speakers [85, 86]. Recently, another speech pathology corpus has been made accessible to the research community: the *Saarbrücken Voice Disorders (SVD)* dataset [87]. This partition contains more than 2000 registers of the vowels /a/, /i/ and /u/ at normal, high, low, and rising-falling pitch; and registers of the sentence “*Guten Morgen, wie geht es Ihnen?*”, recorded at 50 kHz. Due to its novelty, compared to *MEEI*, not many papers have reported results with this corpus.

The *privative* databases exhibit a large variety of characteristics respecting the acoustic conditions followed during the recording process, the instrumentation, the type of speech material that is elicited, the type of disorders, etc. In this regard, most of the *privative* datasets contain microphone recordings of the phonation of the sustained vowel /a/ [52, 88–107] or a combination of several vowels [108–112]. There are some databases with registers of running speech for different languages, which in a text-dependent scenario, employ the repetition of digits [88, 113], the reading of phrases such as “*the rainbow passage*” [114–116], “*the north wind and the sun*” [117, 118], “*the story of Arthur the rat*” [119, 120], or other texts [121–126]. In the text-independent case they employ conversational speech or other types of elicitation tasks [124]. There exist some other databases that include other type of complementary biosignals besides the acoustic material. Namely, some are composed of *Electroglotography (EGG)* recordings of the vowel /i/ [127, 128], /a/ [92, 106]; while others complement the acoustic recordings with questionnaire data [129, 130] or laryngoscopy [130].

7.2 PRE-PROCESSING

Most of the literature employs short-time analysis of speech, as it allows the usage of techniques relying on stationarity assumptions which otherwise could not be used. The short-time analysis is accomplished by means of a framing and windowing procedure, which varies the length or type of the window depending on the applications and of the characterisation techniques. In this regards, the windowing procedure is mostly based on Hamming [32, 38, 39, 45, 85, 86, 89, 97, 103, 113, 120, 121, 129, 131–160] or Hanning [51, 128, 157, 161–169] windows.

Popular preprocessing techniques are based on voice/unvoiced and end-point detectors [45, 85, 106, 120, 162, 165], which ensure that only speech utterances generated during the vibration of the vocal folds are employed; or silence detectors [170] that eliminate utterances not containing speech. Likewise, and to remove the influences of the vocal tract in the speech signal, inverse filtering is often employed [93, 109, 137, 171–175]. Other techniques

include the use of pre-emphasis filtering to accentuate the high frequency content of speech [113, 129, 148], although it has been reported that its use does not improve detection results [176].

7.3 CHARACTERISATION

7.3.1 Temporal and acoustical analysis

Some basic approaches are based on tracking f_0 and deriving high order statistics to track disturbances in the normal vibration patterns of the vocal folds, while some others are based on tracking formants to characterise disturbances in the vocal tract. Some relevant works using such approaches include: [43, 99, 102, 142, 149, 163, 177–179]. These features (particularly f_1 and f_2) have been also employed to determine the jaw-tongue dynamics and provide a distinction between normophonic and pathological voices in [174]. A different approach consists on measuring the vocal function through the quantification of the level or the energy contained in the signals, i.e., the *loudness* [180]. Since this quantity is dependent on the distance between mouth and microphone, measures of *sound pressure level* are preferred instead. These can be achieved by means of intraoral or subglottal pressure apparatuses, or indirectly computed by using accelerometers placed on the neck [116, 181]. Measurements of the vocal level on diverse frequency ranges, leads to a *phonetogram*. This defines voice range profiles which have been used for voice pathology detection [182, 183]. An extension to the method relies on the characterisation of a speech range profile, that has also been used for detection of voice pathologies [117, 184].

Finally, it is possible to extract features from the glottal signal, characterising the opening and closing phases of the glottal waveform as in [93], or via the residue obtained after inverse filtering, with measures such as the *mean square residue* or the *excess coefficient* (kurtosis of the magnitude distribution) [109]. Other acoustic characteristics measure the number and degree of voice breaks and unvoiced frames in speech [185, 186].

7.3.2 Perturbation and fluctuation analysis

Perturbation analysis is referred to the measurement of minor disturbances in the vibration patterns of vocal folds. Two main perturbation effects are identified, one referred to disturbances in amplitude and frequency due to irregular vibration, also called *modulation noise*, and other referred to the effects of noise due to incomplete closure of the vocal folds, also called *additive noise*. With respect to modulation noise, the most prominent family of parameters is related to *jitter*, which measures frequency perturbations among different glottal periods; and *shimmer*, which computes disturbances in amplitude among different glottal periods. Shimmer is one of the most classic parameters, having being widely employed in literature as reported in [93, 125, 149, 163, 177–179, 185–191]. Similarly, jitter has been extensively used,

with some relevant examples in [93, 94, 106, 109, 125, 149, 163, 177–179, 185–191]. However, jitter requires of a precise f_0 estimation, and thus its validity is often questioned [192]. Some methods have been devised to calculate jitter using spectral techniques, avoiding the need of a precise f_0 computation, having accomplished favourable results in pathology detection tasks [164, 166]. Another approach to derive shimmer and jitter based on autoregressive decomposition and pole tracking has been presented in [193]. Regarding the characterisation of additive noise, most of the features measure the relationship between the harmonics and background noise contained in speech. To this end, the *Signal-to-Noise Ratio* (SNR) computes the ratio between signal and noise amplitude [122, 187, 188, 190]; the *Harmonics-to-Noise Ratio* (HNR) measures the ratio between harmonics and noise [125, 149, 163, 186, 188]; *Normalized Noise Energy* (NNE) computes the noise-to-harmonics ratio [84, 178, 188] next to its variation adaptive-NNE [155]; and *Glottal-to-Noise Excitation Ratio* (GNE) measures the correlation between Hilbert envelopes in different frequency bands [157, 194].

By contrast, *fluctuation analysis* is referred to the study of severe disturbances in the dynamics of the vocal folds behaviour, reflecting the inherent instability of the system [192]. *Tremor* is one prominent characteristic that is often studied, referred to low-frequency fluctuation in amplitude and/or frequency, related to pathologies of neurologic origin [192]. Some popular tremor estimates include *amplitude and frequency tremor* [179], or the *turbulent noise index* [98, 140].

7.3.3 Spectral and cepstral analysis

Many systems found in the literature employ spectral or cepstral analysis of speech due to the simplicity they offer, and the ability to characterize both running speech and sustained vowels. Spectral measures derived directly from the speech spectrum include diverse *Long-Time Average Spectrum* (LTAS) characteristics [38, 39, 115, 125], where the spectral tilt [39] -which indicates the degree of energy concentrated in low- vs. high-frequency areas of the spectrum- is a noteworthy example. Other popular sets of features are based on *Linear Prediction Coding Coefficients* (LPC) [144, 195–199], which can also be used to decompose the speech signal into its residual and vocal tract components and hence derive parameters for characterisation [199]. Notable parameters extracted from the residual signal include the *pitch amplitude* [38, 39, 84], which measures the dominant peak in the auto-correlation function of the residual signal; and the *spectral flatness ratio* [38, 84, 109], which measures the residual flatness. Some variations of LPC include cepstral transformation like the *Linear Prediction Cepstral Coefficients* (LPCC) [138, 197, 198], or a mel-transformation of LPC called *Mel-line spectral frequencies* [170].

Other approaches rely on filter-banks to decompose signals into different sub-bands. For instance in [88], correlation functions measure the relationship between the bands of an octave filter-bank for the detection of pathologies. Another example is in [134], where HNR is computed on different frequency bands along with the energy at some critical bands. It is also possible

to employ filter-banks that rely on psychoacoustic criteria to condense spectral information into perceptually-relevant frequency bands. Within this category, *Mel-frequency Cepstral Coefficients* (MFCC) are probably the most popular features [85, 97, 120, 132, 138, 148, 152, 159, 162, 170, 197, 200–202], being often used in combination with its first derivative (Δ) representing velocity [121, 168, 203], and/or its second derivative ($\Delta\Delta$) representing acceleration [133, 146, 147, 158, 165]. Works related to MFCC characterisation include the study of the influence of the tapering window that is employed for the spectrum estimation during the MFCC processing [131], or the derivation of an index based on MFCC and *Gaussian Mixture Model* (GMM)-likelihood metrics: the *pathological likelihood index* [204]. Other psychoacoustic characteristics include *Perceptual Linear Prediction coefficients* (PLP) [120] and its bandpass filtered variation RASTA-PLP [197]; human factor cepstral coefficients [139], or the *band power decorrelation time* obtained through the Meddis and O'Mard filterbank model [45]. The filter-banks can also be employed to perform time-frequency decompositions, through wavelets transformations [91, 97, 100, 140, 141, 154, 195, 205–209] -which are often accompanied by the calculation of energy and/or entropies of the sub-band decompositions- or adaptive time-frequency distribution [210]. A review about these time-frequency decompositions is presented in [211].

Finally, and to characterise additive noise components, measures based on the cepstral prominence of the harmonics are often utilised, by means of features such as the *Cepstral Peak Prominence* (CPP) and its smoothed variation, *Smoothed Cepstral Peak Prominence* (CPPS). A complementary feature is the *Low-to-High Frequency Spectral Energy Ratio* (LHr), which measures the spectral tilt of the spectrum above and below 4 kHz. Relevant works employing these measures include [32, 42, 115, 119, 125, 187, 212]. A derived index that incorporates the information provided by CPP, LHr, and standard deviation of LHr is the *cepstral spectral index of dysphonia* [114].

7.3.4 Complexity features

Complexity is commonly measured through estimators based on nonlinear dynamics. In this regard, the most popular indices compute the dimensionality of the reconstructed space, for discrimination of pathological and normophonic states, by means of the fractal dimension [112] or the *Correlation Dimension* (D_2) [96, 107, 187, 189–191]. Other measures are based on the rate of divergence of trajectories in the phase space. It is expected that the behaviour of normophonic voices differ from those pathological. This has been explored with the computation of *Largest Lyapunov Exponent* (LLE) [94] or the Lyapunov spectrum [213]. Measures of entropy (the rate of information gain) have also been employed in AVQA systems, by means of the first and/or second order entropies [108, 190, 191], the *relative entropy* [101], pseudo-estimators such as the *Ziv-Lempel complexity* [214], or *Hidden Markov Model* (HMM)-based entropy metrics [201, 215]. A concept related to entropy is the *regularity*, which measures the predictability of the time series. The most popular regularity estimator used in pathological voice discrimination

is the *Approximate Entropy* (ApEn) [127, 128, 167, 214]. Other ApEn-derived metrics include the *Sample Entropy* (SampEn) [92], *Gaussian Kernel Sample Entropy* (GSampEn) and *Fuzzy Entropy* (FuzzyEn) [111, 201, 215].

Time-frequency decomposition have been employed to explore the fractal properties of speech [216, 217] or to characterise complexity (using ApEn) on each decomposed sub-band [206]. Other measures explore the use of nonlinear prediction [110, 218]; measure the self-similarity of the voice by means of the *Detrended Fluctuations Analysis* (DFA) [37]; or characterise properties of recurrence to compute the effects of modulation noise using *Recurrence Period Density Entropy* (RPDE)[37].

7.3.5 3-dimensional representations

A popular approach that has been gathering attention lately is based on the multidimensional representations of speech, and the employment of image processing or bidimensional tools for the extraction of characteristics. In this regard, one popular approach is based on *Modulation Spectra* (MS), which characterises the modulation and frequency components of speech. MS produces a 3-dimensional representation that has been employed for the discrimination [136, 159] or assessment of voice pathologies [219, 220].

Following the idea of 3-dimensional representations, authors in [120] define a matrix on the basis of indices of similarity among frames, accompanied by morphological image processing tools for extraction of relevant characteristics. In [221], a mel-spectrogram is characterised by means of the *recurrence texture plots* and the *local binary pattern operator*. Matrices can also be formed using time-frequency decomposition as in [151], where features are extracted through the employment of a multidirectional regression, or the usage of *interlaced derivative patterns* of the glottal source excitation signal [137]. Similarly, authors in [222] extracts features from co-occurrence matrices formed after using filter-banks on the input speech.

7.3.6 Other types of features

There exist some other approaches that do not fit into the above taxonomy. For instance, the multidimensional *acoustic voice quality index* is a metric based on weighted multivariate regression of 6 temporal, spectral and cepstral characteristics that has been used for voice pathology assessment and detection [223]. The use of variograms and the characterisation with the signal-to-dysperiodicity ratio has been explored in [42, 133]. Likewise, decompositions of speech based on non-negative matrix factorisation [210] or empirical mode decomposition [224, 225] have also been reported. The hoarseness diagram has been proposed to visualise additive and modulation noise components [51, 52, 226]. The utilisation of higher order statistics for the characterisation of dysphonic voices is discussed in [156]. In [227], the spectral properties of centralized auto-correntropy is used to detect and classify vocal pathologies.

Some multidimensional approaches consider a combination of several features of different type, having some notable examples in [38, 95, 105, 161, 169, 172, 173, 201, 228, 229]. Other multidimensional approaches consider measures derived from the MPEG-7 standard as described in [146, 230]. Another type of characterisations are based on biomechanical models to describe the behaviour of the glottal and mucosal waveforms [173, 228]. Finally, literature also reports the employment of MFCC features for the construction of a phonological model of 14 features (voicing, place of articulation, turbulence, nasality, etc.) [126].

7.4 DIMENSIONALITY REDUCTION

Dimensionality reduction is aimed to decrease the size of the feature space in order to remove redundant or irrelevant features that might affect performance. Two major types of techniques can be defined: those based on feature extraction, which employ a transformation of the input space; or those based on feature selection, not relying on any transformation. Feature extraction techniques include the classical approaches of *singular value decomposition* [154], *Linear Discriminant Analysis* (LDA) [141, 195, 205, 213, 228, 231, 232], and *Principal Component Analysis* (PCA) [51, 89, 97, 141, 142, 145, 173, 186, 199, 218, 228, 229, 231, 232], or extensions of PCA such as kernel-PCA [145, 229], neural-networks PCA [145], or dynamic feature extraction using PCA [168]. Other type of transformations include those based on HMM [89] or clustering based feature weighting methods [154, 232].

Within feature selection, two types of methods arise. In one hand, the *wrapper feature selection* ties the selection of features to the maximisation of a performance metric obtained with a classifier/regressor. Some notable examples include the use of genetic algorithms to select the best set of features for recognition purposes [91, 97, 130, 206, 213], binary logistic regression analyses with stepwise variable selection [117], sequential backward and forward feature selection [105, 231] or angle modulated differential evolution [103]. On the other hand, *filter feature selection* employs correlation and information approaches to find the most pertaining sets of features. In this respect, literature reports the use of the mutual information [51, 172], correlation analysis [51], Fisher discrimination ratio [147, 214, 221, 228, 230, 233], Fisher discriminant analysis [161], or the Davies-Bouldin index [206].

7.5 MACHINE LEARNING AND DECISION MAKING

The most widely employed decision machines reported in literature include the *Support Vector Machines* (SVM) and GMM classifier. In this respect, SVM has been employed in [88, 91, 93, 100, 102, 103, 111, 126, 130, 133, 136, 137, 141, 153, 159, 160, 163, 168–170, 180, 186, 195, 197, 199, 206, 213, 214, 217, 221, 230, 231, 233, 234]. Similarly, the use of GMM is reported in [131, 138, 144, 147, 151, 158, 159, 168, 170, 196, 197, 200, 201], next to variations such as *Adapted Model coming from an Universal Background Model* (GMM-UBM) [121,

235], *Gaussian Mixture Models - Support Vector Machine* (GMM-SVM) [148, 152, 197, 235] or *i-Vector Modelling* (IV) [235]. Other popular decision machines include *Artificial Neural Networks* (ANN) [90, 97, 108, 141, 156, 162, 186, 187, 202, 205, 207, 213, 231], random forest [103, 105, 153, 188], LDA [39, 42, 84, 140, 149, 158, 170, 179, 208, 209, 225] quadratic-LDA [37], HMM [89, 93, 111, 132, 138, 139, 158, 197, 215], *k-Nearest Neighbour* (KNN) [134, 169, 206] and the Bayes classifier [168] The use of regression techniques is reported in [38, 114, 117, 125, 223].

7.6 SOME RELEVANT WORKS IN AUTOMATIC VOICE QUALITY ANALYSIS SYSTEMS

The present section introduces some relevant works found in the state of the art regarding detection, identification and/or assessment of voice pathologies. For the sake of clarity, they are grouped according to the types of features that have been employed for characterisation.

7.6.1 Temporal-Perturbation

A methodology based on 21 perturbation features, for the characterization of the recordings of the MEEI database (53 control and 638 pathological), is presented in [140]. To identify the glottal cycles that are necessary for the calculation of perturbation parameters, a cross-correlation detector is used. Classification is carried out using a LDA in a leave-one-out scheme, obtaining a classification accuracy of 96%. In [157], a set of perturbation features is also tested, with special emphasis on the GNE parameter. Experiments performed in the MEEI database confirm the discriminant ability of GNE, achieving a 90% detection accuracy with no cross-validation of the results. The performance of a telephone-based voice pathology system is investigated in [149]. Experimentation is performed in a subset of the MEEI database (573 pathological and 58 normophonic), by means of perturbation features and a LDA classifier. The recordings are distorted to simulate telephone transmission. The achieved classification accuracy within this approach is of 74%, inferior to the one obtained in a controlled environment (89%). When grouping disorders according to some aetiological criteria, an accuracy of 87% is obtained identifying neuromuscular disorders, 78% physical abnormalities and 61% in a mixed pathology scenario. To determine whether a correlation exists between the GRBAS scale and the *Multidimensional Voice Program* (MDVP) parameters, 37 patients reading the *rainbow passage* are perceptually evaluated by means of the GRBAS scale, and acoustically analysed using the MDVP [236]. Using a multivariate regression model, a significant correlation between 3 perturbation parameters of MDVP and the components of the GRBAS scale is identified. Authors in [38] describe a study aiming to determine the relationship between acoustic measures and auditory-perceptual dimensions of overall voice severity and pleasantness. To this end, recordings of 30 dysphonic and 6 control speakers reading the *rainbow passage* are employed.

Non-trained listeners assess the degree of dysphonia according to the pleasantness of the uttered speech. Acoustic measures based in [LTAS](#) are used to train a regressor, and to assess the predictive potential of the acoustic parameters in determining the perceptual severity and pleasantness. Results indicate that the acoustic measures accounted for 48% of overall voice severity and 40% of voice pleasantness for dysphonic speakers. By fusing the acoustic information of the features and the auditory-perceptual ratings, a classification accuracy of 100% is obtained in distinguishing between normophonic and dysphonic voices.

A method for measuring jitter, which is relatively devoid of f_0 computation, is explored in [\[164\]](#). The method, called *spectral jitter*, is tested out in two databases: [MEEI](#) (53 normophonic and 657 pathological) and *Hospital Universitario Príncipe de Asturias* ([HUPA](#)) (238 normophonic and 201 dysphonic), and compared to the jitter estimates provided by two programs ([MDVP](#) and Praat). Results suggest that the spectral jitter achieves improvements up to 4% in absolute performance, compared to the other estimators in pathological voice discrimination tasks. Spectral jitter is further tested in [\[166\]](#), using a short-time analysis scheme applied to running speech and sustained vowels of the same databases. In a cross-database experiment, the operating point that maximizes *Area Under Receiver Operating Characteristic Curve* ([AUC](#)) in [MEEI](#) is employed in posterior experiments with the other database. Therefore, the threshold that provides an [AUC](#) = 0.95 in sustained vowels of [MEEI](#) is employed in the [HUPA](#) partition, obtaining an [AUC](#) = 0.68, whereas in the running speech partition of [MEEI](#) it is found that [AUC](#) = 0.88.

Finally, in [\[231\]](#), 22 long-time acoustic features extracted using the [MDVP](#) are evaluated. 2 feature extraction and 4 feature selection techniques are used to find the set of features providing the best accuracy. In the same manner 6 classifiers, including a [SVM](#), are employed for detection of pathologies in the [MEEI](#) database. Results indicate that [LDA](#) for feature extraction and [SVM](#) for classification, provide the best performance, with a recognition rate of 94%. Following the same idea, authors in [\[233\]](#) employ the 22 parameters of [MDVP](#) for detection and classification of voice pathologies. Experiments are performed on three pathologies: vocal-folds cyst, unilateral vocal-fold paralysis, and vocal-fold polyps; using recordings of the sustained phonation of vowel /a/ of [MEEI](#), an Arabic voice pathology database and a subset of the [SVD](#). Using the Fisher discrimination ratio to rank the parameters and a t-test between normophonic and pathological populations, accuracies of 99%, 88%, and 72% are obtained for [SVD](#), [MEEI](#), and the Arabic voice pathology database respectively.

7.6.2 Spectral and cepstral analysis

[LPC](#) are used to calculate the vocal tract area from speech signals. Then, the variance, skewness, and kurtosis of this area is calculated along time and post-processed via [PCA](#) for dimensionality reduction. Tests are made in two databases containing the sustained phonation of vowel /a/: the [MEEI](#) database (partition as in [\[84\]](#)), and a subset of the [SVD](#) database (1342 patho-

logical and 686 normophonic speakers). Using a SVM for classification the system achieved a 99% of accuracy in the MEEI database and 95% in the SVD dataset. Additionally, in a one-vs.-all identification of pathologies task (considering only vocal fold cyst, paralysis and polyps), results ranged from 89% to 94% in MEEI and from 94% to 99% in the SVD database.

MFCC features and their first and second derivatives are extracted from a subset of the MEEI database containing 125 recordings (53 normophonic and 82 pathological) of the sustained vowel /a/. Using a *learning vector quantizer* a performance of 96% is obtained. A refinement of the previous system is presented in [147], replacing the former classifier by a GMM and using the subset of patients in MEEI that is recommended in [84]. Additionally, a dimensionality procedure based on the f_β and Fisher's discriminant ratio is employed to investigate the influence of the derivatives in the detection task. The authors report an accuracy of 94% in the extended set of the dataset, indicating that the inclusion of the second derivative provides no further improvement to the accuracy of the system. Similarly, MFCC features and the same MEEI partition are also employed in [85]. There, several methodological concerns regarding the design of systems to automatically detect pathologies are discussed. With the proposed methodology, aimed to permit comparisons with other studies in the literature, a performance up to 90% is obtained by means of ANN. Likewise, MFCC features are utilized in [162] to characterise voiced segments extracted from text-dependent running speech. Taking registers of the vowel /a/ as a baseline acoustic material and text recordings of the *rainbow passage* (both coming from the MEEI database), a classification accuracy of 93.8% is attained by means of an ANN in the sustained vowels dataset, whereas a 96.3% is obtained using the voiced segments extracted from running speech.

In [152] MFCC and GMM-SVM are used for the discrimination of pathological and normophonic voices in the MEEI database (partition as in [84]). To this end, an *Universal Background Model* (UBM) is built by concatenating the whole database and then generating GMM-UBM models through *Maximum a-posterior* (MAP) adaptation. Classification is carried out by means of linear sequential kernels, providing an accuracy of 96%, superior to the one attained using a GMM (91%). In a similar work, GMM-SVM are employed but using Kullback-Leibler and Bhattacharyya distances kernel [203]. Experiments are performed using MFCC features and its derivatives, plus normalized energy parameters, in a subset of the SVD database containing 40 female voices suffering from spasmodic dysphonia and 60 control patients uttering the /a/ vowel. Accuracy of 96% is attained in a scenario with no cross-validation. Authors in [121], used running speech to detect and assess the grade of pathology in a database composed of 60 dysphonic and 20 normophonic patients reading a passage. By means of GMM-UBM models adapted through MAP techniques, applied to an auxiliary UBM dataset of normophonic speakers, a classification accuracy of 85% is obtained. In the assessment task results are about 60% with most of the errors produced in the intermediate classes of the grading decision.

Working on the idea of exploiting the sex of the speaker for pathology detection, authors in [202] employ MFCC and ANN for identification purposes. Studies effectuated in the MEEI database (partition as in [84]) demonstrate that voice pathology detection errors decrease (from 12% to 11% in absolute terms) when the sex of the speaker is considered in the system.

A methodology based on *wavelet packet decomposition* and *local discriminatory bases* is discussed in [208]. The aim is to analyse a subset of the MEEI database (51 normophonic and 161 pathological) using LDA and features extracted from the local discriminatory bases. A performance of 96% is obtained in the normophonic/pathological classification task; and of 74% in a four-group classification of male-normal, female-normal, male-pathological and female-pathological. Similarly, features extracted from a 1-D discrete wavelet decomposition and LPC modelling the residual signal of wavelet sub-bands are investigated in [195]. In tests applied to a subset of the MEEI database (53 normophonic and 240 pathological), a classification accuracy of 97% is obtained with no cross-validation and SVM. A time-frequency approach is presented in [209]. Features extracted from a time-frequency decomposition of speech are employed to train a LDA classifier. Tests performed in a subset of the MEEI database containing recordings of the *rainbow passage* (51 normophonic and 161 pathological), report an accuracy of 93%.

In [136], MS features are employed along SVM for three discrimination tasks: normophonic vs. pathological, classification of four voice disorders (vocal polyps, adductor spasmodic dysphonia, keratosis, vocal nodules), and classification of paralysis vs. the four disorders combined. Using the MEEI database (with the partition in [84]) and a SVM, an accuracy of 94% is obtained in normophonic vs. pathological detection, of 85% in paralysis vs. non-paralysis, whereas accuracy ranged from 87% to 92% in the identification task. In a similar way, authors in [159] perform experiments in the MEEI database (partition as in [84]) and the HUPA database (199 normophonic and 200 pathological). The authors utilize MFCC features to train a GMM and MS features to train a SVM. Then, fusing the scores of both classifiers, the maximum attained accuracy in HUPA is of 84%, whereas it is of 96% in MEEI.

In [219], a system to automatically assess the $\mathcal{Q}$ and $\mathcal{R}$ traits, according to the GRBAS perceptual scale is presented. Experiments are performed on the MEEI database after a perceptual analysis of the whole dataset given by 3 evaluators. Using a new set of features based on morphological properties of the MS, as well as GMM for classification and PCA and LDA for dimension reduction, an accuracy of 81% is obtained for detecting $\mathcal{Q}$, and 84% for detecting $\mathcal{R}$. In comparison, MFCC and their derivatives achieved 80% and 77% in the detection of $\mathcal{Q}$ and $\mathcal{R}$ respectively. This scheme is refined in [220], where the features presented above are further studied and optimized to maximize performance. Experiments are effectuated in three datasets: MEEI, HUPA and *Hospital Gregorio Marañón* (GMar) databases. Classification rates up to 90% are obtained in MEEI, 71.1% in HUPA and 72% in GMar. The severity of dysphonia in pre and post-therapy is evaluated in [36]. 208 female patients with muscle tension dysphonia reading the *rainbow passage* are assessed by experts, judging the severity of the disease in a 100-points scale, in both pre

and post conditions. Using CPP, some of its moments, and 2 features derived from spectral analysis, the authors found that the proposed feature set provides strong correlations with the perceived dysphonia severity rating.

Spectral moments derived from LTAS (spectral mean, standard deviation, skewness, and kurtosis) and CPP measures are employed in [115], for assessing voice quality in a subset of the MEEI database composed of 27 subjects with voice disorders and 27 control speakers. The voice severity is assessed using the CAPE-V scale. Results indicate that the 2 cepstral and 3 LTAS measures present moderate to strong correlations to overall voice severity and serve to differentiate between normophonic and dysphonic speakers. An auditory filter-bank approach for the calculation of LTAS features is proposed in [45], to qualitatively identify the most useful spectral characteristics for voice pathology detection. By using sustained vowels of the HUPA database, and running speech and sustained vowels of the MEEI dataset, the authors report minor differences in the power spectral envelope of normophonic and dysphonic voices of sustained vowels. However, and according to a new measure called *decorrelation time*, it is found that power in bands 2 kHz to 6.4 kHz is significantly less stable in dysphonic voices. Using running speech, significant differences in the power level of frequency bands above 5.3 kHz are reported between control and dysphonic voices.

7.6.3 Complexity

One of the complexity features that has been studied the most in AVQA systems is the D_2 . For instance, authors in [107] use it to characterize signals according to their typology, in a database composed of 122 pathological registers of the vowel /a/. Results indicate that D_2 reveals significant differences among all types of signals. Posteriorly, authors in [189] combined D_2 with jitter and shimmer to analyse sustained vowels of normophonic and unilateral laryngeal paralysis patients in a subset of the MEEI database (57 control and 67 pathological). Jitter and shimmer are calculated only for nearly periodic voices, obtaining significant differences between pathological and normophonic classes ($AUC=0.7$ for jitter and $AUC=0.8$ for shimmer). D_2 is calculated for all voices obtaining significant differences between normophonic and pathological states ($AUC=0.74$). Authors suggest that perturbation and complexity features might characterize voice signals due to their complementarity. A further investigation, including D_2 , jitter, shimmer *signal-to-noise ratio* and second-order entropies, is presented in [190]. To analyse sustained phonation, the vowel /a/, and vowels extracted from the reading texts of the MEEI database are employed. Results revealed the ability of complexity features to discriminate between pathological and normophonic cases. On contrary, perturbation measures performed well with sustained vowel, but should be used with caution in the vowels extracted from running speech.

Two new complexity features are introduced in [37], for detection of voice disorders: RPDE and DFA. Using *Quadratic discriminant analysis* and bootstrap resampling for validation, a classification accuracy of 92% is obtained, using the proposed features solely in the MEEI database. This result is higher than

the one obtained combining other classical features, that do not exceed 81%. New features are also proposed in [201], where two entropy measures based on Markov models are introduced: *Rényi HMM Entropy* (**rHMMEn**) and *Shannon HMM Entropy* (**sHMMEn**). These new features are utilized to characterize the **MEEI** database (partition as in [84]), along with other 9 complexity features, 3 perturbation and **MFCC** parameters. Following a score-level classifier fusion, that combined the perturbation and **MFCC** parameters with the complexity analysis characteristics, an accuracy of 98% is reached using a **GMM**.

In [237], D_2 is employed to characterize changes between pre and post-treatment in a dataset composed of 88 female speakers suffering from muscle tension dysphonia. The resulting features are correlated with the ratings given by 10 listeners which quantify the severity of dysphonia using a 100-points scale. Results show a significant difference between pre and post treatment voice samples. However, the authors found that one fourth of the dysphonic voices could not be analysed by the inability of D_2 to converge to a value due to the severity of the impairment. In addition, D_2 is poorly correlated with perceived dysphonia severity ratings as a measure of pre/post-treatment changes in moderate to severe dysphonia, resulting in weak correlations with listener-perceived severity and poor utility as an index of treatment change. Similarly, authors in [118], used 3 complexity measures (**DFA**, **RPDE** and D_2) and 3 perturbation features (jitter, shimmer and the inverse of **HNR**) to assess the severity of voices between pre and post states. Experiments are performed in a dataset of 11 control and 17 patients with unilateral vocal fold paralysis, assessed by means of the **GRBAS** scale before and after a surgery to correct the disorder. Results indicate that only jitter and shimmer are sensitive to the changes in voice quality after surgery.

7.6.4 Multiples sets of features

This subsection includes those works which encompass more than one set of characteristics. For instance, authors in [169] employ 11 sets of features to characterise spectral, perturbation and cepstral aspects of the voices for detection of pathologies and identification of three classes (normophonic vs. nodular vs. diffuse). The sustained phonation of the vowel /a/ is recorded three times from 25 healthy subjects, 25 with diffuse disorders, and 54 with nodular impairments. Classification is carried out by means of a **KNN**, a **SVM**, and a committee of classifiers. The best classification accuracy that is obtained for the detection case is of 95.1%, and of 84.6% for the identification case. Authors also found that the perturbation measures are the most informative.

Following a different approach, a set of 87 temporal and spectral features are extracted from the **MEEI** database in [161]. However, instead of using the features for training a classifier, the degree of correlation between the features is employed for classification purposes. The resulting correlation matrix (of size 3741) is first preprocessed by means of feature extraction and selection procedures. Then, a decision is taken by thresholding the most relevant features. By using two spectral characteristics and their correla-

tion, a correct classification rate of 94.7% for pathological voices and 89.5% for normophonic is found. Authors in [145], extract a set of 28 perturbation, spectral, acoustic features (jitter, shimmer, energy, formants, MFCC, etc.) from a subset of the SVD database composed of 1342 speakers uttering the /a/, /i/ and /u/ vowels at high, low and normal pitch, and having 4 diseases (hyperfunctional dysphonia, functional dysphonia, laryngitis and vocal folds paralysis). The methodology includes a transformation based on PCA and two variations of PCA (built on kernel and ANN methods). The analysis considers: each vowel individually at different intonations, men and women separately for each pathology and, each vowel individually at different pitches. The results in the distinct combination of tests per sex, disease and vowel varies in the ranges of 86% to 100%.

Following a multi-modal approach, authors in [130], combine colour images of the vocal folds, the phonation of the sustained vowel /a/ and questionnaire data provided by patients to detect pathologies. To characterize images of the vocal folds, 8 feature sets are employed, whereas 14 questions responded by patients constitutes the query dataset. In addition 12 feature sets of different types (perturbation, spectral, cepstral, etc.) extracted from voice are employed. A SVM classifier carries out classification tasks, whereas the selection of the best feature sets and the optimization of the SVM parameters is embedded into a single process based on a genetic search. Experiments are performed in a database composed of 151 patients with nodular lesions, 64 with diffuse disorders and 25 controls. The classification accuracy reached 90% for the image modality, 88% for voice and 85% for the query data. However, accuracy increases to 98% when combining the different sources of data. Continuing with the idea of fusing multi-modal information, authors in [129] present an approach that combines features extracted from acoustic sources and query data for the design of a pathology detector. 26 features (intensity, loudness, MFCC, f_0 , etc.), their deltas and moments are extracted from a database containing 459 control and 903 pathological speakers uttering the vowel /a/. The authors found that the query modality outperformed voice, but its fusion, by means of a random forest, produces the best results with an accuracy of 95%.

7.7 DISCUSSION

The present section has introduced some of the techniques that literature reports for the design of AVQA systems. Those were described using the taxonomy introduced in Figure 5.1. Based on the information extracted from literature, this thesis will consider different sets of features for characterisation purposes, including some measuring the effects of perturbation, spectral or cepstral characteristics of speech, modulation spectra and complexity features. In the same manner, the classification is to be carried out using GMM, one of the most popular classifiers found in the state of the art. Moreover, some regression techniques and ordinal classification models are to be employed for the purposes of voice pathology assessment. Despite they have not been -to the author's knowledge- employed in literature, they

fit perfectly with the ordinal nature of the assessment task, and provide the possibility of extending ordinal decisions to a continuous space. Due to the documented problems of the [MEEI](#) database, it has been preferred to employ the [SVD](#) partition (which is publicly available) and other two private databases belonging to the research group for the majority of the tests, whereas other datasets are used as ancillary speech material. More details about the techniques that are used, the databases, and the methodological stages followed are presented in the following chapters.

EXTRACTION OF RELEVANT PARAMETERS

As illustrated in Figure 5.1, the first compounding block of an AVQA scheme is referred to the extraction of relevant parameters from the speech signal, which is composed of the following stages: (i) *preprocessing*, where operations to adequate the input signal are followed, through normalisation, windowing and/or framing procedures; (ii) *characterisation*, where features aiming at describing voice production phenomena are extracted; (iii) *dimensionality reduction*, where extraction or selection procedures transform, select or rank the most relevant characteristics for the detection/identification/assessment task. Each one of these stages is to be explained next.

8.1 PREPROCESSING

Since speech is intrinsically non-stationary, some preprocessing procedures are employed before the utilisation of conventional signal processing techniques which often rely on stationarity requirements. One common procedure is the *short-time analysis*, which decomposes the input signal into a series of equal-length chunks of speech, called *frames*, permitting the treatment of each individual chunk as a stationary or quasi-stationary fragment. This procedure is composed of two operations: *framing*, which divides the signal into frames (typically overlapped); and *windowing*, which tapers the beginning and ending of the frames, through the product with a window function to improve their spectral properties. In this way, a speech signal $s[t]$ is short-time analysed by multiplying it with a sliding tapering window function $w[t]$ of length W , such that F frames $s_f[t]$ are obtained [238]:

$$\{s_f[t]\}_{f=1}^F := \sum_{m=-\infty}^{\infty} s[m]w[f-m] \quad (8.1)$$

The window function $w[f-m]$, should be selected to provide a frequency response with a narrow bandwidth in the main lobe and large attenuation in the side-lobes. One popular choice fitting such requirements is the Hamming window which is as follows:

$$w[t] = \begin{cases} 0.54 + 0.46 \cos\left(\frac{t\pi}{W}\right), & 0 \leq t \leq W \\ 1, & \text{otherwise} \end{cases}$$

The selection of the window length, W , is of great importance. For applications using sustained phonation the duration of the window is typically

set in between 20-40 ms. The upper limit ensures that frames are not that large to make the quasi-stationarity assumption void. The lower limit is set to make the analysis independent of the location of pitch pulses within the segment, and to include at least two to three times pitch periods. Since the typical range of pitch frequency is between 80 – 500 Hz, a pitch pulse is expected every 12 – 2 ms [239]. For applications using running speech, windows lengths are typically set in the order of 20 – 30 ms to conserve the quasi-stationary assumptions [80], and to ensure that phonemes which are of the order of 80 ms are correctly characterized within the processed frame.

8.2 CHARACTERISATION

The characterisation stage has the goal of extracting features capable of portraying the properties of normophonic and dysphonic conditions. The idea is to extract from each windowed frame, a d -dimensional vector of characteristics, $\vec{x} = \{x[1], \dots, x[d]\}$, describing properties of the segments, and which is associated to a label ℓ , indicating the membership of the segment.

A commonly utilised descriptor of vocal condition is the level of **vocal aperiodicity**, as, phenomenologically, it is related to diverse aspects of voice production. Three types of vibrational patterns are often encountered in voice signals, including [192]: (i) *Type I* voices, characterised by a nearly periodic behaviour; (ii) *Type II* signals, which contain bifurcations, sub-harmonics or modulating frequencies; and (iii) *Type III* voices, which are characterised by an aperiodic behaviour. In accordance with such distinction, normophonic voices are usually enclosed into the Type I typology, whereas pathological voices are embodied into the Type II and III categories [39, 192]. As a matter of example, Figure 8.1 illustrates some cases of voice signals following the above-mentioned typology.

Figure 8.1. Taxonomy of voice signals according to [192]. *Top panel*: normophonic type I signal, characterised by a periodic behaviour; *middle panel*: pathological type II signal having modulating frequencies; *bottom panel*: pathological type III signal characterised by its aperiodic behaviour.

Vocal aperiodicity is explained as the result of some very distinctive processes occurring during the voice production process such as [42, 240, 241]: (i) *irregular dynamics* of the vocal folds and involuntary transients between dynamic regimes; (ii) *modulation noise* owing to extrinsic perturbations in amplitude and frequency of the glottal cycle; and (iii) *additive noise* owing to turbulent airflow. A dysphonic voice is said to possess one or more of these aperiodicity features. Indeed, modulation noise is often associated to the presence of the perceptual characteristics of roughness, while additive noise is associated to breathiness, two trademarks of voice pathology [164]. By contrast, irregular dynamics are distinguishing features encountered on very specific voice impairments such as *diplophonia* or *biphonation*, and as such **are not considered in the remaining of this thesis**.

Above-mentioned facts suggest that an analysis of the level of vocal aperiodicity may be valuable when designing AVQA systems. Notwithstanding, finding characteristics that effectively describe the presence of modulation and additive noise for classification/assessment purposes is difficult, specially since some degree of aperiodicity is present even in non-pathological states due to perturbations inherent to the phonation process [216]. As a result, there is no single feature, in the context of screening, that perfectly differentiates between normophonia and pathology, or that biunivocally correlates acoustic measurements and voice quality [204]. A common approach to counteract this, consists on studying different acoustic features, in the hope of finding combinations of characteristics that complement with each other. In this sense, the best characteristics would be those with the lowest correlation with the others but capable to provide the best discrimination capabilities [157]. Although multidimensional studies have reported good performance in screening tasks, this type of analysis is usually carried out by complex pattern recognition techniques, which makes difficult the interpretation of results from the perspective of a human evaluator [157].

With respect to the type of features employed in AVQA systems, literature reports diverse characterisation schemes which have been found to perform differently according to the pathologies under study or the database that has been employed. In general, some popular sets include *perturbation*, *spectral-cepstral*, *modulation spectra* or *complexity* features. These sets of characteristics are to be explained in the next sections.

8.2.1 Perturbations parameters

Perturbations are minor disturbances or temporary changes that deviate from an expected behaviour [192]. Perturbation parameters have been frequently used to analyse vocal aperiodicities that are produced by the effect of modulation or additive noise. In this regard, the most popular modulation noise quantifiers includes families of parameters based on the estimation of *jitter* and *shimmer*. On one hand, jitter is a measure of the short-term (cycle-to-cycle) perturbation of the vocal f_0 . Jitter parameters include [242]:

- *Jitter relative*: average absolute difference between consecutive periods, divided by the average period.
- *Jitter Relative Average Perturbation (RAP)*: average absolute difference between a period and the average of this and its two neighbours, divided by the average period.
- *Jitter Five-point Period Perturbation Quotient (PPQ5)*: average absolute difference between a period and the average of this and its four closest neighbours, divided by the average period.

On the other hand, shimmer measures short-term (cycle-to-cycle) estimator of amplitude perturbations. Shimmer parameters include [242]:

- *Shimmer absolute*: variability of the peak-to-peak amplitude in decibels.
- *Shimmer relative*: average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude
- *Shimmer Three-point Amplitude Perturbation Quotient (APQ3)*: average absolute difference between the amplitude of a period and the average of the amplitudes of this and its neighbours, divided by the average amplitude.

Despite both families of features have been of considerable utility for describing type I signals, its validity for type II and III typologies has been put into question due to the need of an accurate identification of f_0 , a difficult task in highly non-stationary dysphonic voices [37, 192, 243]. For this reason other metrics not depending on f_0 estimation are often preferred. In this respect, the most popular perturbation measures quantifying the additive noise content of the signal include *HNR*, *NNE* and *GNE*, which have being extensively applied in the evaluation of voice quality correlating with many speech disorders and perceptual ratings [45, 157].

8.2.1.1 Harmonics-to-Noise Ratio

HNR is a widely popular estimator of the additive noise [244]. This feature relies on computing separately the harmonic and the noise components of speech. The harmonic structure (ι_{En}) is simply quantified, for each frame, as the average energy of the speech signal. To compute the noise component, the input signal is divided into chunks of speech according to the pitch period. Then, the noise (γ_{En}), is measured as the mean energy of the differences between the energy of individual chunks and ι_{En} . Finally, *HNR* is the log-ratio between ι_{En} and γ_{En} :

$$\text{HNR} = 20 \log \frac{\iota_{En}}{\gamma_{En}} \quad (8.2)$$

Some shortcomings of the above methodology include the need of a correct identification of f_0 , and the assumption that the harmonic structure of speech is maintained along time irrespective of modulation noise effects.

With this in mind, a further improvement based on cepstral domain techniques is proposed in [245], with a measure called *Cepstral Harmonics-to-Noise Ratio* (CHNR). The idea of CHNR is to consider the noise component as the energy that cannot be related to the spectrum of the original signal, and to calculate HNR as the difference in level between the cepstral total energy and the noise energy [194, 245]. To compute CHNR, the following steps are followed: (i) the cepstrum of the signal is calculated; (ii) then, the harmonic components of the signal are identified as the periodic *rahmonics* (equivalent to harmonics in the cepstrum); (iii) these are removed by a liftering operation to obtain the equivalent noise energy; (iv) finally, the level difference between the previously calculated noise energy and the total cepstral energy defines CHNR. A graphic depicting the algorithm for the calculation of CHNR is presented in Figure 8.2. Moreover, the speech spectrum of a normophonic voice signal and the noise calculated using the CHNR method is introduced in Figure 8.3. As observed, the noise component of the dysphonic voice is larger than that of the normophonic one.

Figure 8.2. Algorithm for the calculation of the CHNR.

Figure 8.3. Magnitude spectrum superimposed to the noise spectrum calculated using the CHNR noise estimation method for a normophonic and pathological voice. CHNR is the difference in level between both components.

8.2.1.2 Normalized Noise energy

NNE is another measure employed to quantify the level of additive noise, and which is defined as the ratio between the energy of the noise (γ_{En}) and total energy of the signal, (ι_{En}) [246]:

$$\text{NNE} = 20 \log \frac{\gamma_{En}}{\iota_{En}} \quad (8.3)$$

As observed, this is simply the inverse of [HNR](#), and similarly needs of a correct estimation of f_0 . Compared to [HNR](#), [NNE](#) varies in the calculation of the ι_{En} term. This is determined differently during the harmonic peaks of the signal or the valleys in-between peaks. The procedure is illustrated in Figure 8.4, and can be summarised as follows: (i) the fundamental frequency of the signal and the log-spectrum are computed; (ii) next, the noise energy in the valleys is calculated directly from the spectrum, whereas the noise energy in the harmonic peaks is estimated by interpolating the minima of adjacent valleys; (iii) finally, [NNE](#) is calculated as the level difference between the spectral total energy and the noise energy. A graphic depicting the speech spectrum of a normophonic voice signal and the noise calculated using the [NNE](#) method is presented in Figure 8.5. As observed, the noise component of the dysphonic voice is larger than that of the normophonic.

Figure 8.4. Algorithm for the calculation of [NNE](#).

Figure 8.5. Magnitude and noise spectrum of a normophonic and pathological voice. The noise spectrum is calculated directly from the valleys, whereas the values in the harmonics part are interpolated from the values of adjacent valleys. [NNE](#) is the level difference between both spectra.

8.2.1.3 Glottal-to-noise excitation ratio

Despite [NNE](#) and [CHNR](#) have been designed to measure the level of additive noise in voices, they are sensitive to the presence of modulation noise. Attempting to counteract this problem, [GNE](#) has been proposed to characterise turbulent noise while disregarding the effects of modulation [194]. [GNE](#) assumes that glottal pulses produce a simultaneous and synchronous excitation of different frequency channels, which is expressed by the correlation between Hilbert envelopes of different frequency bands [51]. Since

the envelopes in all channels share the same shape, a high correlation between them is expected [51, 194]. However, the presence of turbulent noise decreases correlation, and measuring this loss of correlation becomes useful to describe the presence of aspiration.

The algorithm to calculate the **GNE** is depicted in Figure 8.6 and described next [51, 194]: (i) input speech is downsampled to 10 kHz; (ii) then, an inverse filtering operation is applied to obtain an estimate of the speech source; (iii) Hilbert envelopes of different frequency bands with a fixed bandwidth and different centre frequencies are calculated; (iv) the cross correlation is then computed for every pair of envelopes for which the difference of their centre frequencies is equal or greater than half the bandwidth; (v) finally, the maximum absolute value among all correlation functions is the **GNE** value. Figure 8.7 depicts the correlation matrix of a normophonic and a dysphonic voice. As observed, there is a decrease in correlation in pathological states.

Figure 8.6. Methodology to estimate the **GNE**.

Figure 8.7. Correlation matrix during the **GNE** computation for a normophonic and dysphonic voice.

8.2.2 Spectral-Cepstral analysis of speech

Measures derived from the acoustic spectrum/cepstrum have been widely used in the study of pathological phonation. Indeed, spectral and cepstral features have demonstrated high correlation with the perceptual assessment of dysphonia, providing large sensitivity when used in classification tasks [32], and high reliability predicting the type and severity of dysphonia [34]. In the same manner, and from the study of breathy voices, it has been found that the presence of turbulence noise is associated with some distinctive

spectral properties [247]: (i) an elevated amplitude spectrum with a high degree of inter-harmonic noise or harmonic substitution by noise; (ii) noise regions of high intensity in the spectrum above the primary formants; (iii) antiformants in the low frequency spectrum due to tracheal resonance; and (iv) high relative intensity in the fundamental harmonic compared to adjacent harmonics, global spectral energy or first formant. Aforesaid facts suggest that spectral and cepstral features might be exploited for the analysis and characterisation of pathological phonation. The analysis with spectral/cepstral features also presents several advantageous properties, including its appropriateness for analysing both sustained vowels and running speech with no extra procedures, and the ability to characterise speech signals without depending of a f_0 estimation [32, 45]. With this in mind, the following section introduces some popular spectrum/cepstrum measures extracted from voice.

8.2.2.1 Cepstral peak prominence

A number of studies involving hypofunctional, hyperfunctional, and mixed-aetiology speakers have highlighted the benefits of using the CPP in clinical settings [32]. These studies report CPP as one of the strongest correlates of breathiness -in both sustained vowels and running speech tasks- with an inverse relationship between the severity of perceived breathy voice quality and the value of CPP [32, 36, 45]. CPP is a normalised measure of the cepstral peak amplitude, which compares the level of harmonic organisation of the speech to the cepstral background noise resulting from aspiration [247, 248]. The rationale behind CPP calculation is that a highly periodic signal should possess a well defined harmonic structure and, consequently, a more prominent cepstral peak than one having a decreased periodic structure. The prominence of the peak (the normalized cepstral peak) is preferred over a measurement of the absolute amplitude because the cepstral peak is affected by the global energy and the window size of the cepstrum analysis. Overall amplitude normalisation can be achieved by fitting a linear regression line relating *quefreny* (equivalent to frequency in the cepstral domain) to the cepstral magnitude. The line is computed between 1 ms and the quefreny range, which is typically set to 60-300 Hz [248]. The difference in amplitude between the cepstral peak and the corresponding value on the regression line directly below the peak (overall signal amplitude) is the CPP [247].

The estimation of the cepstral peak might be improved using a smoothed version of the cepstrum as proposed in [248]. The resulting method is called CPPS and can be described in the following way: (i) windowed frames of speech are log-cepstrum transformed; (ii) a time-quefreny liftering operation is applied to smooth the log-cepstrum; (iii) the overall amplitude is then estimated by means of a fitting procedure that calculates the first order polynomial of the cepstrum in the quefreny range 60-300 Hz; (iv) the quefreny that produces the largest peak is evaluated in the fitted polynomial; (v) the largest peak value is then subtracted from the previously fitted polynomial evaluation, thus obtaining the CPPS. A schematic diagram summarising the

CPPS calculation process is shown in Figure 8.8. Additionally, a graphic depicting the behaviour of CPPS in a normophonic and a pathological signal is presented in Figure 8.9.

Figure 8.8. Depiction of the CPPS calculation process

Figure 8.9. CPPS for a normophonic and a pathological voice.

8.2.2.2 Low-to-High frequency spectral energy ratio

Several studies have indicated that cepstral measures may be supplemented by other acoustic quantifiers such as the LHR to account for larger degrees of variance between perceived and acoustically predicted severity ratings [36]. LHR is simply the ratio between the average spectral energy below 4 kHz to the average energy above 4 kHz.

8.2.2.3 Mel-frequency Cepstral Coefficients

MFCC are a popular set of features in speech-related applications, that exploit auditory principles and the decorrelating property of the cepstrum to characterize speech signals. MFCC are based on psychoacoustic studies that indicate a nonlinear relationship between the real frequency of a sound and the actual frequency that is perceived. One scale measuring such relationship is described in [249], and is termed *mel scale*. In there, authors report experiments in which participants are initially presented with a 1 kHz tone and are then instructed to adjust the tone so that their perceived frequency is twice that of the initial frequency. Next, they are instructed to adjust to tone to be perceived as ten times that of the original frequency, half of the initial frequency and so on. The graphic that is achieved from plotting the results of the experiments defines a mapping between the real and the perceived fre-

quency, which is approximately linear below 1 kHz and logarithmic above [238], and which is expressed mathematically as [250]:

$$F^{mel} = \frac{1000}{\log[2]} \log \left[1 + \frac{F^{Hz}}{1000} \right]$$

where F^{mel} is the frequency in the mel scale and F^{Hz} is the frequency in Hz.

Several reasons justify the utilisation of MFCC in speaker recognition. For instance, calculation of MFCC do not require pitch detection. It is also robust against some types of voice distortion, and compresses information on the first part of the cepstrum, hence providing some dimensionality reduction [202]. In voice pathology applications, MFCC have demonstrated an inherent ability to model either the irregular movement of the vocal folds or its lack of closure. Indeed, the alterations related with the vocal fold vibration (modulation noise) are reflected in the low bands of the MFCC, whereas the higher bands model turbulence components (additive noise) [147].

MFCC can be estimated using a *Discrete Time Fourier Transform* (DTFT) via a non-parametric FFT approach. Following a short-time analysis, MFCC are calculated as the *Discrete Cosine Transform* (DCT) of the logarithmic energy in several frequency bands as follows [147]:

$$\{\text{MFCC}\}_{m=1}^{N_{MFCC}} = \sum_{r=1}^{N_{fil}} \log \mathfrak{J}_r \cos \left(\frac{m\pi(r-0.5)}{N_{fil}} \right) \quad (8.4)$$

where N_{MFCC} is the order of the MFCC coefficients; N_{fil} is the number of mel bands in the mel scale; and $\mathfrak{J}_r$ is:

$$\mathfrak{J}_r = \sum_{j=1}^{N_{FFT}} w_r^{mel}[j] |S[j]|$$

where w_r^{mel} is the triangular function associated with the r -th mel band; $|S[j]|$ is the magnitude spectrum estimated via FFT; and N_{FFT} is the number of FFT components. The filters remove the harmonics of the spectrum, providing spectral smoothing and making an estimate of the spectral envelope. The filtering is designed to provide a high spectral resolution at lower frequency bands, performing better than a set of regularly spaced filters. The triangular filter-bank that is typically employed is depicted in Figure 8.10. In addition, a graphic summarizing the MFCC computation is shown in Figure 8.11. Finally, Figure 8.12 introduces a representation of a normophonic and a pathological voice, where each point in the horizontal axis of the graphic is referred to a windowed frame of speech.

8.2.2.4 Perceptual linear predictive coefficients

PLP have been designed in accordance to a scale modelling the human auditory system [251, 252]. The procedure aims at replicating the perceptual mechanisms of hearing to model the vocal tract transfer function. To this end, three considerations are taken. First, the Bark scale is employed to represent,

Figure 8.10. Depiction of the mel-filterbank

Figure 8.11. Depiction of the MFCC calculation process

as in the the mel-scale case, the psycho-acoustical differences between perceived pitch and the actual frequency, by simulating the cochlear response of hearing. An approximation of the Bark frequency relating frequency in Hz (F^{Hz}) and in Barks (F^{Bark}) is described as [80]:

$$F^{Bark} = 13 \arctan \left(0.00076 F^{Hz} \right) + 3.5 \arctan \left(\frac{F^{Hz}}{7500} \right)^2$$

Second, since the auditory system is more sensitive to frequencies ranging from 100 Hz to 6 kHz (with the greatest sensitivity around 3-4 kHz) an equal loudness-level curve is used to equalise the dissimilarities in perceived loudness [11]. Third, an intensity-to-loudness conversion is applied. This is an approximation to the power law of hearing and simulates the nonlinear relationship between the intensity of sound and its perceived loudness [252]. The procedure for obtaining PLP coefficients is presented in Figure 8.13 and is summarised as follows: (i) the power spectrum of windowed frames of speech is calculated; (ii) this is filtered by means of a Bark filter-band, which is of the form shown in Figure 8.14; (iii) a pre-emphasis operation is then

Figure 8.12. MFCC of a normophonic and a pathological signal.

employed to simulate the frequency sensitivity of human hearing; (iv) to approximate the nonlinear relationship between intensity and perceived loudness a cubic root amplitude compression is applied; (v) finally, coefficients are obtained by means of **LPC** modelling. A depiction of **PLP** for a normophonic and dysphonic voice is presented in Figure 8.15. Each point in the horizontal axis of the graphic corresponds to a windowed frame of speech.

Figure 8.13. Depiction of the **PLP** calculation process

Figure 8.14. Depiction of the Bark filterbank

Figure 8.15. **PLP** of a normophonic and a pathological signal.

8.2.3 Modulation Spectra features

Despite **MS** is a spectral technique, it is treated differently in this thesis due to its multi-dimensional character. **MS** is a three-dimensional representation, where abscissa corresponds to the *modulation frequency*, ordinate axis to the *acoustic frequency*, and applicate to the *acoustic energy*. This representation allows observing different voice features simultaneously such as the harmonic nature of the signal and the frequency modulation of f_0 and harmonics [219]. To obtain **MS**, the input signal is filtered by means of a **sTFT** filter-bank (as in Equation (1.2)). Then, a second **sTFT** computes the modulation frequencies

of the sub-bands of the previous output. The result of following such procedure is a matrix $\mathbf{E}_{(A \times M)}$, composed of A acoustic and M modulation bands. Since $\mathbf{E}$ contains real and imaginary parts, the modulus of the spectrum is represented as $|\mathbf{E}|$. A value at any point can be described as $\mathbf{E}(f^a, f^m)$, where f^a is the central frequency at the acoustic band, and f^m is the central frequency of the modulation band. The computation of the matrix $\mathbf{E}$ using the MS approach is presented in Figure 8.16.

Figure 8.16. Stages followed to compute the MS matrix.

After obtaining the matrix $\mathbf{E}$, parameters are extracted to characterise the properties of pathological and normophonic voices. To this end, the features employed in [220] are described in the next paragraphs. In addition, Figure 8.17 summarises the computation procedures of these features.

Figure 8.17. Stages followed to compute the MS features.

8.2.3.1 Modulation Spectrum Homogeneity

The matrix $\mathbf{E}$ can be portrayed as a two dimensional image, allowing a visual inspection of its properties. Following such approach, authors in [220] observed that pathological voices usually contain irregular distributions, in opposition to those of normophonic signals which seemed to be more homogeneous and possess less contrast. One example of this behaviour for a

normophonic and dysphonic voice is presented in Figure 8.18. To quantify the level of homogeneity in modulation spectra representations *Modulation Spectra Homogeneity* (MSH) has been proposed, computing its value with a variation of the method in [253], which is as follows:

$$\text{MSH} = \sum_a \sum_m \left(E(f^a, f^m) - \overline{E(f^a, f^m)}_{N_s \times N_s} \right), \quad (8.5)$$

where $\overline{E(f^a, f^m)}_{N_s \times N_s}$ is the average in a $N_s \times N_s$ window centred at f^a, f^m .

Figure 8.18. MS modulus of a (a) normophonic and (b) dysphonic voice.

8.2.3.2 Cumulative Intersection Point and Rate of Points above Linear Average

Since MS differs in normophonic and pathological conditions, a histogram calculated on the MS modulus is expected to reflect the effects of voice disorders. In particular, authors in [220] state that pathological voices usually have a large number of points with energy above the average value of $\overline{E}$, and that this represent an energy dispersion in the central modulation band (0 Hz) towards the side bands. Having this in mind, two metrics are proposed to measure that dispersion: *Cumulative Intersection Point* (CIL) and *Rate of Points Above Linear Average* (RALA).

CIL is the intersection that results when plotting two curves: the histogram-increasing cumulative curve and the histogram-decreasing cumulative curve. As a matter of illustration, the CIL metric calculated from a normophonic and dysphonic voice is presented in Figure 8.19. It is important to remark that the histograms are processed from the MS modulus in logarithmic units (dB).

On the other hand, RALA is defined as the number of points in the MS modulus which are above $|\overline{E}|$ (in linear units), divided by the number of points which are below this average, and which is defined as follows:

$$\text{RALA} = \frac{\sum_{f^a} \sum_{f^b} \Omega(f^a, f^m)}{N_T - \sum_{f^a} \sum_{f^b} \Omega(f^a, f^m)} \quad (8.6)$$

where N_T is the total number of points and:

$$\Omega(f^a, f^m) = \begin{cases} 1, & |E(f^a, f^m)| \geq |\overline{E}| \\ 0, & \text{otherwise} \end{cases}$$

Figure 8.19. CIL, calculated as the cumulative ascending and descending histograms for a (a) normophonic and (b) dysphonic voice.

Figure 8.20 represents these points in a normophonic and a pathological voice. It is noticeable that, as expected, the MS of dysphonic voices present more points above the modulus average.

Figure 8.20. Representation of points above $\overline{|E|}$ for the computation of RALA for a (a) normophonic and (b) dysphonic voice.

8.2.3.3 Modulation Spectrum Percentile

Some statistical measurements called *Modulation Spectrum Percentile* (MSP) are also calculated from the linear histograms. These are the 25, 75, and 95 percentiles extracted from the MS modulus (MSP_{25} , MSP_{75} and MSP_{95} respectively), serving as indicators of the distribution of levels at low, mid-low, and high level ranges.

8.2.4 Complexity analysis

Complexity is a controversial term that has been classically linked to randomness and mistakenly associated to information measures such the algorithmic complexity [254]. However, it is more appropriately related to a "meaningful structural richness" [255] or to *fractal* behaviour rather than to randomness. Complex behaviour is typically observed in biological systems that manifest at least one of the following dynamical properties [256]: (i) nonlinearity; (ii) nonstationarity; (iii) time irreversibility (or asymmetry); or (iv) multiscale variability. With respect to the pathology detection/assess-

ment task, the interest in using complexity analysis arises from the theory of complexity loss in ageing and disease or *decomplexification theory*. This theory postulates that healthy systems possess intrinsically complex variability resulting from long range correlations and nonlinear interactions, but the level of complexity decreases with ageing and disease progression [257]. Consequently, a quantification of this complexity loss might be employed as an indicator of disease and hence utilised for the design of AVQA systems. With this in mind, the following section introduces some concepts related to complexity analysis based on *Nonlinear Dynamics Analysis* (NDA), long-term correlations descriptors and entropy quantifiers.

8.2.4.1 Characterization using complexity analysis

The most straightforward approach to investigate the nonlinear dynamical properties of a system is through NDA. It is well known that nonlinear phenomena arises naturally in physiological systems, and voice production is not an exception to this regard. Indeed, supporting findings of nonlinearity in phonation include nonlinear pressure-flow relations in the glottis, the delayed feedback from mucosal wave, the nonlinear stress-strain curves of vocal fold tissues, nonlinearities associated with vocal fold collision [258], or asymmetries between the right and left vocal folds vibrations [259]. In addition, nonlinear dynamical behaviour of models of the vocal folds such as period-doubling (subharmonics), bifurcations, and transitions to irregular vibration have been observed in experiments with excised larynges. Likewise, period-doubling bifurcations and chaos-like features have been observed in signals from patients with organic and functional dysphonia [37]. Aforementioned facts suggests the appropriateness of using NDA to characterise the dynamics of voice production even in pathological scenarios, since voice pathologies can be considered disorders of glottal dynamics [260].

The usual approximation to NDA requires of a reconstruction stage to reveal the system dynamics. To this end, the input time series is described using a d_e -dimensional space termed *state space* populated by d_e -dimensional points called *state vectors*. The temporal evolution of these state vectors conform a trajectory that characterizes the dynamical behaviour of the system. If on average this trajectory contracts -such as in the case of dissipative systems- a closed subset of the state space, called *attractor*, is formed.

One difficulty encountered in the representation stage comes from the impossibility of accessing to the actual system's variables but to a time series, which is simply an observable of the system dynamics. To deal with such situation, the *time-delay embedding theorem* is usually employed [261]. This ensures the reconstruction of an equivalent attractor holding the same topological properties of the original attractor of the system, by means of a procedure termed *embedding*. Thus, given a time series $\{s[t]\}_{t=1}^T$, the reconstructed (or embedded) state vectors $\vec{s}_t$ are obtained as follows:

$$\vec{s}[t_i] = \{s[t], s[t - \tau], \dots, s[t - d_w]\} \quad (8.7)$$

where the value $d_w = (d_e - 1)\tau$ is the *embedding window* defining the span of analysis of the current state vector within the time series; d_e is the *embed-*

Figure 8.21. Three-dimensional projection of an attractor embedded using uniform embedding for a (a) normophonic and (b) dysphonic voice; and non-uniform embedding for a (c) normophonic and (d) dysphonic voice.

ding dimension or the smallest number of required coordinates to represent the state vectors without overlapping in the state space; and τ is a constant called *time lag* (or *time delay*) influencing the shape and spreading of the reconstructed attractor. Since τ takes a constant value, the reconstruction procedure is also termed *uniform embedding*. As a matter of example, Figures 8.21a and 8.21b illustrate an uniform embedding reconstruction for a normophonic and pathological voice respectively.

The uniform embedding technique has been widely used for reconstructing attractors, providing successful results in several pattern recognition tasks. Nonetheless, it suffers from issues when dealing with non-stationary biosignals with multiple periodicities. In fact, when choosing τ , a short time lag might be optimal for high frequency components, whereas a long lag might perform better for low frequency components and modulations. Therefore, a compromise might be inadequate for both time-scales [262]. Bearing that in mind, it has been discussed that a *non-uniform embedding* approach replacing the single time delay τ with a lag vector $\vec{\zeta} = [\zeta_1, \zeta_2, \dots, \zeta_{d_w}]$ might account for such divergences in time scale [262].

The literature presents some procedures for searching an optimum lag vector $\vec{\zeta}^{opt}$ in a non-uniform embedding reconstruction scheme. For instance, an early approach is based on the *Minimum Description Length* (MDL) principle

[262], which relates the quality of the embedding to the quality of modelling. Another approach is grounded on the *False Nearest Neighbours* (FNN) criterion, using the ratio of false neighbours for increasingly adding time delays until this ratio drops to zero [263]. Similarly, a feed-forward neural network trained by structural learning is suggested for finding the optimal vector in [264]. In [265], two statistics are introduced for evaluating the quality of a non-uniform embedding reconstruction. A geometrical criterion for selecting the optimal time lag vector is also described in [266], aiming to relate the magnitude of the attractor's spreading in the state space to the reconstruction quality. This approach is further improved using evolutionary algorithms as discussed in [267]. A final scheme is presented in [268], by means of an objective function guiding the search of an optimal state-space reconstruction in terms of noise amplification and minimization of the complexity of reconstruction.

One common characteristic of all the above referred approaches is that they do not provide a closed procedure to find the best delay vector. Besides, they present a high computational cost, since in most cases it is necessary to model and optimize parameters, or to employ greedy search algorithms to find the optimal $\vec{\zeta}^{opt}$. Therefore, they are unsuitable for pattern recognition tasks. To overcome these issues providing a proper non-uniform embedding reconstruction within the framework of the nonlinear characterization of biosignals, a method that attempts to simultaneously minimize the redundancy and maximize the relevance of reconstructed vectors in a feature selection-like schema is presented in [269]. To this end, a lag vector is initially set to $\vec{\zeta} = [\zeta_1, \zeta_2, \dots, \zeta_{d_w^{max}}]$, with $\zeta_1 = 1$, $\zeta_2 = \zeta_1 + 1$, $\zeta_3 = \zeta_2 + 1$, $\dots$, and the maximum embedding window limit set to $d_w^{max} = (d_e - 1)\tau$. The use of such initialization permits to consider all the possible state vectors $\{\vec{s}_t\}$ within a window d_w^{max} , such that an expanded embedding matrix constituting the complete space to be reduced can be represented as follows:

$$M = \begin{bmatrix} s[t] & s[t-1] & \dots & s[t-d_w^{max}] \\ s[t-1] & s[t-2] & \dots & s[t-(d_w^{max}+1)] \\ \vdots & \vdots & \ddots & \vdots \\ s[t-(T-d_w^{max})] & s[t-(T-d_w^{max}+1)] & \dots & s[t-T] \end{bmatrix}$$

The expanded matrix rewritten in terms of its columns is as follows:

$$M = [\vec{\eta}^1, \dots, \vec{\eta}^i, \dots, \vec{\eta}^{d_w^{max}}]$$

where each column $\vec{\eta}^i$ in M is related to the element ζ_i in $\vec{\zeta}$. By using the expanded matrix, the task of finding an optimal lag vector is transformed into other aiming to find the most informative set of columns $\vec{\eta}^i$ in M . The solution provided in [269] relies on following an unsupervised feature selection scheme. In order to reduce the complexity of the searching procedure, the dimensionality of the embedding matrix is fixed in advance with the value given by d_e using the FNN method under the assumption that it is robust enough to clearly distinguish the dynamics of the system. The feature

selection problem is carried out following an irrelevance and redundancy analysis, through the usage of an unsupervised heuristic algorithm termed *Unsupervised Minimal Redundancy/Maximal Relevance (UmRMR)* [270]. Figures 8.21c and 8.21d illustrate the resulting non-uniform embedding reconstructions for a normophonic and dysphonic voice respectively.

Once the state space is reconstructed it is possible to extract features that describe the dynamics of the embedded attractor. Classical NDA characteristics include descriptors of the dimensionality of the reconstruction, such as D_2 or LLE. In the context of voice pathology analysis, another feature called RPDE is also typically employed. However, the extraction of classical NDA features relies on the assumption that the dynamics of the system are purely deterministic [201]. This cannot always be satisfied since randomness is an inherent part of speech production. Certainly, phonation is produced as a combination of deterministic components owing to the non-linear movements of the vocal folds tissues and bulk of the air in the vocal tract, plus a stochastic component due to the high frequency aeroacoustic pressure fluctuations caused by vortex shedding at the top of the vocal folds [37]. Additionally, stochastic effects arise in breathy voices due to aspiration noise, for which it might be infeasible to completely characterize the speech dynamics by using NDA characteristics solely, specially for severely disordered voices with dominant stochastic components. An alternative to quantifying the degree of complexity of a dynamical system which does not depend on the deterministic requirement, include features intended as entropy estimators of the reconstructed time series, or long-range correlation measures extracted directly from the time series.

With respect to the entropy estimators, they rely on the computation of the *entropy* of a system, which is the rate of generation of new information. For a random variable $\mathcal{A}$ with alphabet $\mathfrak{A}$ and probability function $p(a) = \Pr\{\mathcal{A} = a\}, a \in \mathfrak{A}$, the *Shannon entropy* is defined as [255]:

$$\mathcal{H}(A) = - \sum_{a \in \mathfrak{A}} p(a) \log p(a) \quad (8.8)$$

Similarly, the *Rényi entropy* of order r_α can be defined as follows:

$$\mathcal{H}^{r_\alpha}(A) = \frac{r_\alpha}{1 + r_\alpha} \sum_{a \in \mathfrak{A}} p(a)^{r_\alpha} \quad (8.9)$$

To quantify the entropy of an embedded system the *Kolmogorov-Sinai entropy* is usually employed. Its computation requires partitioning an attractor into n_c hypercubes of radius ϵ and content ϵ^{d_e} , measuring the states of the system at infinitesimal intervals of time Δt . For this reason, a close solution is not feasible since it will require extremely long noise-free time series which are hardly obtainable. A family of statistics measuring regularity-related but not intended as Kolmogorov-Sinai estimators of entropy- has been proposed [255]. The use of such features has been explored in [201], where in combination with perturbation and cepstral/spectral features have improved the performance of an AVQA system. Some commonly used regularity features include the *ApEn*, *SampEn*, *Modified Sample Entropy (mSampEn)*,

GSampEn and FuzzyEn. Other type of quantifiers which might also be employed to quantify an attractor's entropy include *Permutation Entropy* (PE), rHMMEn and sHMMEn.

There also exist some metrics which are devoted to characterise long-range correlations of the time series, aiming to capture the *multiscale* or *fractal* properties of the systems. In this regard, a fractal is complex-like object, which may be generated by stochastic or nonlinear deterministic mechanisms, composed of subunits (and sub-subunits) that resemble the larger scale structure. This property is known as *self-similarity* [257]. The idea of using long-range correlation measures comes from indications suggesting fractal properties in both normal and disordered speech due to the presence of turbulence [37], as evidenced theoretically and experimentally [271]. Some common long range quantifiers include *Hurst Exponent* (h^e) and DFA.

All the features presented in this section are to be briefly described in the next paragraphs.

8.2.4.2 Correlation dimension

D_2 quantifies the auto-similarity of an embedded attractor [272], being usually employed as an indicator to differentiate determinism from stochastic behaviour. Systems whose dynamics are governed by stochastic processes are said to possess an infinite value of D_2 because random processes fill very large-dimensional subsets of the system state space. By contrast, a finite non-integer value of the dimension is considered to be an indicator of determinism [273]. Despite this belief, it has been shown that stochastic processes provide finite and predictable values for D_2 [273, 274], and thus this measure should be used with care. The Takens' estimator, $\mathcal{J}(\epsilon)$, is usually employed to calculate D_2 , relying on the computation of the correlation sum:

$$c_{\Sigma}(\epsilon) = \lim_{T \rightarrow \infty} \frac{1}{T^2} \sum_{i,j=1; i \neq j}^T \mathfrak{d}^H(\vec{s}[t_i], \vec{s}[t_j]; \epsilon)$$

where $\vec{s}[\cdot]$ are the reconstructed state vectors, ϵ is a threshold, and $\mathfrak{d}^H$ is a distance function defined in terms of the Heaviside function ϑ :

$$\mathfrak{d}^H(\vec{s}[t_i], \vec{s}[t_j]; \epsilon) = \vartheta(\epsilon - \|\vec{s}[t_i] - \vec{s}[t_j]\|)$$

From all the distances between state vectors, a sample of size N_p is chosen, assuming that they are independent and randomly distributed according to:

$$P(\mathfrak{d}^H(\vec{s}[t_p], \vec{s}[t_q]; \epsilon) \mid_{p,q=1}^{N_p} < \epsilon) = c_{\Sigma}(\epsilon) = A \cdot \epsilon^{D_2}$$

where A is a constant. Then, the Taken's estimator is defined as:

$$\mathcal{J}(\epsilon) = \left[\frac{-1}{N_p - 1} \sum_{p=1}^{N_p} \log \left(\frac{\mathfrak{d}^H(\vec{s}[t_i], \vec{s}[t_j]; \epsilon)}{\epsilon} \right) \right]^{-1}$$

In the limit $\epsilon \rightarrow 0$ and $N_p \rightarrow \infty$, $\mathcal{J}(\cdot)$ is the estimator of D_2 [275]. In practice, the value of D_2 corresponds to the fitted line inside the straight region of the curve $\mathcal{J}(\epsilon)$ vs. $\log(\epsilon)$. An example of the the estimation of D_2 for a normophonic and dysphonic voice, and a range of embedding dimensions, is presented in Figure 8.22.

Figure 8.22. Estimation of D_2 as the linear zone of the curve for a range of d_e , for a normophonic and a pathological voice.

8.2.4.3 Largest Lyapunov Exponent

LLE measures the divergence of nearby orbits in state space. **LLE** can be estimated as the slope of the function $Q(\Delta t)$ (in a region exhibiting a robust linear increase) within a range Δt . The function $Q(\Delta t)$ is as follows [272]:

$$Q(\Delta t) = \frac{1}{T_0} \sum_{t_0=1}^{T_0} \log \left(\frac{1}{|\mathcal{U}(\vec{s}[t_0]; \epsilon)|} \sum_{\vec{s}[t_i] \in \mathcal{U}(\vec{s}[t_0]; \epsilon)} |\vec{s}[t_0 + \Delta t] - \vec{s}[t_i + \Delta t]| \right)$$

where $\vec{s}[t_0]$ is some initial condition in an appropriate embedding space and $\mathcal{U}(\vec{s}[t_0]; \epsilon)$ is the neighbourhood of $\vec{s}[t_0]$ enclosed by a ball of radius ϵ . If a negative slope is found, then **LLE** would suggest fixed point attractors, whereas zero values would represent limit cycles and positive values noisy signals. A graphic illustrating the computation of **LLE**, as the slope between $Q(\Delta t)$ and Δt , is presented in Figure 8.23.

Figure 8.23. Estimation of **LLE** as the slope of the curve $Q(\Delta t)$ vs. Δt , for a normophonic and dysphonic voice.

8.2.4.4 Recurrence Period Density Entropy

RPDE is proposed in [37], as a method to quantify the effects of modulation noise in speech. It is based on the method of close returns, measuring statistics from recurrent trajectories that converge to a certain region of state space. **RPDE** is computed as follows: for a certain reference point $\vec{s}[t_0]$, a neighbourhood $\mathcal{U}(\vec{s}[t_0], \epsilon)$ is defined by a ball of radius ϵ centred at $\vec{s}[t_0]$.

The trajectory of $\vec{s}[t_0]$ is followed forward in time until it leaves the ball, and the time it takes to first return to the ball is computed. This time difference, called recurrence time Δt_r , is then computed for all embedded data points. Next, a histogram of recurrence, $R(\Delta t_r)$, is constructed and normalized to obtain a recurrence-time probability density such that:

$$p(\Delta t_r) = \frac{R(\Delta t_r)}{\sum_{i=1}^{T_{\max}} R(\Delta t_i)} \quad (8.10)$$

where $T_{\max}$ is the maximum recurrence time in the attractor.

Finally, **RPDE** is defined as the Shannon entropy (Equation (8.8)) of $p(\Delta t_r)$ normalized in the interval $[0, 1]$ and which is as follows:

$$\text{RPDE} = \frac{-\sum_{i=1}^{T_{\max}} p(\Delta t_i) \log p(\Delta t_i)}{\log T_{\max}} \quad (8.11)$$

A graphic illustrating the recurrence-time histograms for the computation of **RPDE** is presented in Figure 8.24 for a normophonic and a dysphonic voice.

Figure 8.24. Recurrence time histograms in a normophonic and pathological voice for the computation of **RPDE**.

8.2.4.5 Hurst Exponent

h^e is an estimator of self-similarity, indicating the persistence of the system to follow trends. It varies between 0 and 1, in such a manner that $h^e = 0$ indicates white noise; $h^e < 0.5$ represent antipersistent behaviour (the current trend might change the direction of tendency in the following iteration); $h^e = 0.5$ indicates lack of correlation between current and past values (typical from Brownian movement); whereas $h^e > 0.5$ indicates persistent behaviour (the current trend of the system tends to be maintained). There are several algorithms for the computation of h^e , but the one employed in this thesis is the so called *absolute values of the aggregated series* [276], which works as follows. First, the time series $s[t]$ is divided into T/L blocks of size L :

$$g_k = \frac{1}{L} \sum_{i=(k-1)L+1}^{kL} s[i]; k = 0, 1, 2, \dots$$

Then, the sum of the absolute value of the aggregated series is computed:

$$\mathfrak{g}_k^{abs} = \frac{1}{T/L} \sum_{k=1}^{T/L} |\mathfrak{g}_k| \quad (8.12)$$

Finally, the logarithmic $\mathfrak{g}_k^{abs}$ at different window sizes L is computed and plotted. If the original time series has long-range dependence with parameter h^e , the result should be a line with slope h^e-1 . A graphic illustrating the computation of h^e , as the line fitted to the $\mathfrak{g}_k^{abs}$ vs. L curve, for a normophonic and dysphonic voice is presented in Figure 8.25.

Figure 8.25. Estimation of h^e using the absolute values of the aggregated series for a normophonic and pathological voice.

8.2.4.6 Detrended Fluctuations Analysis

DFA is an estimator of self-similarity that serves to compute the stochastic components in speech due to aspiration noise [37]. The value of DFA gives hints about the trends followed by the system. In this manner, if $DFA < 0.5$ the signal is uncorrelated (characteristic of white noise); if $DFA = 0.5$, the signal is anti-correlated (negative trends); if $DFA > 0.5$, there are positive correlations (positive trends) [277]. The first step to estimate DFA consists on integrating the time series to create a new stochastic process that exhibits self-similarity over a large range of time scales as [278]:

$$\eta_p = \sum_{t=1}^p (s[t] - \bar{s})$$

where $p = 1, 2, \dots, T$; and $\bar{s}$ is the mean of the time series. Then, η_p is divided into windows of length L . Within each window, a line is fit using a least square procedure. This represent the *local trend* of the window or η_p^L . The *Root Mean Square Error (RMSE)* between the integrated time series and the local trends is:

$$\mathfrak{B}(L) = \sqrt{\frac{1}{T} \sum_{p=1}^T (\eta_p - \eta_p^L)^2} \quad (8.13)$$

For a broad range of window sizes, it expected that $\mathfrak{B}(L) \propto L^{DFA}$. Finally, DFA is then computed after plotting $\mathfrak{B}(L)$ vs. $\log(L)$, and finding the slope of the resulting line [37]. A graphic illustrating the computation of DFA for a normophonic and dysphonic voice is presented in Figure 8.26

Figure 8.26. Estimation of DFA as the slope of $\mathfrak{B}(L)$ vs. $\log L$ for a normophonic and pathological signal. The slope of the graph is the estimated DFA value.

8.2.4.7 Regularity quantifiers

ApEn computes the *logarithmic likelihood* that in a time series of length T , patterns of size $d_m + 1$ are within tolerance r of each other, given that patterns of size d_m are already within tolerance r [274, 279, 280]. ApEn can be extracted directly from embedded attractors or from the time series. In this manner, and when extracted from attractors, d_m equates d_e ; otherwise, it is simply set to a certain value (e.g., $d_m = 2$). The ApEn statistic is defined as:

$$\text{ApEn} = \phi^{d_m}(r) - \phi^{d_m+1}(r) \quad (8.14)$$

where:

$$\begin{aligned} \phi^{d_m}(r) &= \frac{1}{T - d_w} \sum_{i=1}^{T-d_w} \log C_i^{d_m}(r) \\ C_i^{d_m}(r) &= \frac{1}{T - d_w} \sum_{j=1}^{T-d_w} \mathfrak{d}^H(\vec{s}[t_i], \vec{s}[t_j]; r) \end{aligned}$$

and the distance measure $\mathfrak{d}^H$ as in Equation (8.2.4.2).

Despite ApEn has been successful in biosignal characterization, several modifications have been proposed to correct a biasing phenomenon due to self-matches [281]. The first modified statistic is the SampEn, described as:

$$\text{SampEn} = -\log \left(\frac{\hat{\phi}^{d_m}(r)}{\hat{\phi}^{d_m+1}(r)} \right) \quad (8.15)$$

where $\hat{\phi}^{d_m}(r)$ is as in Equation (8.14) but excluding self-matches.

It has been stated that ApEn and SampEn have validity and precision problems for being formulated on the non-continuous Heaviside function [282]. A first attempt to overcome this issue is proposed in [282] with the *Gaussian Kernel Approximate Entropy* (GApEn) which replaces the Heaviside function in Equation (8.14) by a Gaussian Kernel function as:

$$\mathfrak{d}^G(\vec{s}[t_i], \vec{s}[t_j]; r) = \exp \left(-\frac{\|\vec{s}[t_i] - \vec{s}[t_j]\|^2}{10r^2} \right)$$

The authors originally formulated GApEn in terms of ApEn, but it is possible to avoid the self-matching phenomena as in SampEn, and define GSampEn.

A second attempt to improve the problems of the Heaviside function is proposed in [283, 284] by means of the **FuzzyEn**. The authors propose to replace the Heaviside function by a Fuzzy membership function of the form:

$$\mathfrak{d}^F(\vec{s}[t_i], \vec{s}[t_j]; r) = \exp \left(- \frac{\|\vec{s}[t_i] - \mathcal{P}_0[t_i] - \vec{s}[t_j] - \mathcal{P}_0[t_j]\|^{\theta_F}}{r} \right)$$

where θ_F is a parameter determining the shape of the fuzzy membership function and $\mathcal{P}_0[t_i]$ removes the baseline of space state vectors as follows:

$$\mathcal{P}_0[t_i] = \frac{1}{d_w} \sum_{j=0}^{d_w-1} \vec{s}[t_{i+j}]$$

A third attempt is proposed in [285] by means the **mSampEn** that employs a simplified sigmoid defined as:

$$\mathfrak{d}^m(\vec{s}[t_i], \vec{s}[t_j]; r) = \frac{1}{1 + \exp \left(\frac{\|\vec{s}[t_i] - \mathcal{P}_0[t_i] - \vec{s}[t_j] - \mathcal{P}_0[t_j]\| - 0.5}{r} \right)}$$

As a matter of example, Figure 8.27 presents the regularity estimators as extracted from a normophonic and a pathological voice.

Figure 8.27. Regularity estimators for a normophonic and a pathological voice. Each point corresponds to a windowed frame.

8.2.4.8 Permutation entropy

PE performs an entropy quantification through the use of permutation patterns calculated from the input time series [286, 287]. This characteristic has received a lot of attention in the last years thanks to its simplicity, robustness and the low computational cost associated to its calculation. To calculate **PE** the following procedure is followed [286, 287]: given a state vector $\vec{s}[t_i]$ of dimension d_e , an ordinal pattern is extracted and each one of the elements in the state vector are sorted in ascending order. Then, a permutation pattern $\vec{s}_\pi[t]$ is created with the offset of the permuted values. As a matter of example take the state vector $\vec{s}[t_0] = [3, 1, 4]$. The vector is firstly sorted in ascending order generating the sequence $[1, 3, 4]$, which results in a permutation pattern $\vec{s}_{\pi_0} = [1, 0, 2]$. In case that two values are equal, they are ordered according to their time of appearance [287], e.g, the vector of values

$[1, 4, 1]$, leads to the permutation $\vec{s}_{\pi_0} = [0, 2, 1]$. With the permutation $\vec{s}_{\pi_i}$ it is possible to derive a measure of probability as:

$$p(\vec{s}_{\pi_i}) = \frac{f^{patt}(\vec{s}_{\pi_i})}{T - d_w}$$

where $f^{patt}(\cdot)$ stands for the frequency of the pattern.

Finally considering all the possible permutations (in total, $d_e!$) the PE is defined as the Shannon Entropy of the probability distribution:

$$\text{PE} = \sum_{i=1}^{d_e!} p(\vec{s}_{\pi_i}) \log p(\vec{s}_{\pi_i}) \quad (8.16)$$

PE ranges from 0 for a monotonously increasing or decreasing time series (determinism), to $\log d_e!$ for white noise (stochastic behaviour). A graphic illustrating PE for the frames of a normophonic and pathological voice is presented in Figure 8.28.

Figure 8.28. PE entropy of a normophonic and a pathological voice. Each point corresponds to a different windowed frame.

8.2.4.9 HMM-based entropies

Entropy estimators calculated on embedded attractors only quantify the divergence of the trajectories of the attractor but do not take into account the directions of divergence. To include this information, two measures based on *Markov Chains* (MC) are proposed in [201, 288]. A MC is a random process that can take a finite number of values k , at certain moments of time ($t_0 < t_1 < t_2 < \dots$). There exist some processes that can be seen like a MC, whose outputs are random variables generated from probability functions associated to each state. Such processes are called *Hidden Markov Process* (HMP), since the states of the Markov process cannot be identified from its output, i.e, the states are *hidden*. A HMP can also be understood as a Markov process with noisy observations. Using a *Discrete Hidden Markov model* (DHMM) to represent a stochastic process, the noise is modelled by means of discrete distributions, and finally, it is possible to obtain a *Probability Mass Function* (pmf) for the noise in each state. Denoting the actual state of the process in time t as $\mathfrak{S}_t$, a DHMM can be characterized by the following parameters:

- $\sigma = \{\sigma_i\}; i = 1, 2, \dots, n_\theta$ is the initial state distribution, where $\sigma_i = p(\mathfrak{S}_0 = i)$ is the probability of starting at the i -th state and n_θ is the number of states in the Markov chain.

- $\mathfrak{Y} = \{\mathfrak{Y}_{ij}\}; 1 \leq i, j \leq n_\theta$, is the set of transition probabilities among states, where $\mathfrak{Y}_{ij} = p(\mathfrak{S}_{t+1} = j \mid \mathfrak{S}_t = i)$, is the probability of reaching the j -th state at time $t + 1$, coming from the i -th state at time t .
- $\mathfrak{D} = \{\mathfrak{D}_{ij}\}; i = 1, 2, \dots, n_\theta; j = 1, 2, \dots, n_v$, is the probability distribution of the observation symbol; n_v is the total number of symbols; and $\mathfrak{D}_{ij} = p(o_t = v_j \mid \mathfrak{S}_t = i)$ where o_t is the output at time t and v_j are the symbols that can be associated to the output.

All parameters are subject to standard stochastic constraints. Using this definition, an empiric entropy, H_e , can be defined as follows:

$$\mathcal{H}_e = \mathcal{H}_{MC} + \mathcal{H}_G$$

where $\mathcal{H}_{MC}$ is the entropy due to the Markov process and $\mathcal{H}_G$ is the entropy due to noise. By using the Shannon entropy (see Equation (8.8)), $\mathcal{H}_E$ can be rewritten as a new entropy measure called **sHMMEn**:

$$\text{sHMMEn} = - \left(\sum_{ij}^{n_\theta} \sigma_i \mathfrak{Y}_{ij} \log \mathfrak{Y}_{ij} + \sum_{i=1}^{n_\theta} \sum_{j=1}^{n_v} \mathfrak{D}_{ij} \log \mathfrak{D}_{ij} \right) \quad (8.17)$$

If a Rényi entropy is used (see Equation (8.9)), **rHMMEn** is defined instead:

$$\text{rHMMEn} = \sum_{i=1}^{n_\theta} \frac{\sigma_i}{1 - r_\alpha} \log \sum_{j=1}^{n_\theta} \mathfrak{Y}_{ij}^{r_\alpha} + \sum_{i=1}^{n_\theta} \frac{1}{1 - r_\alpha} \log \sum_{j=1}^{n_v} \mathfrak{D}_{ij}^{r_\alpha} \quad (8.18)$$

where r_α is the entropy order ($r_\alpha > 0, r_\alpha \neq 1$) usually set to 2.

A graphic illustrating the computation of **rHMMEn** and **sHMMEn** for a normophonic and a pathological voice is portrayed in Figure 8.29.

Figure 8.29. Computation of **rHMMEn** and **sHMMEn** for a normophonic and dysphonic voice. Each point represents a windowed frame.

8.3 DIMENSIONALITY REDUCTION

The characterisation stage is often accompanied by a *dimensionality reduction* procedure, specially if the feature space is too large. The objective is to reduce the number of variables of the original space, in the search of a decrease in the computational burden, an improvement in the accuracy of the analyses, and the avoidance of problems such as the curse of dimensionality [289]. Dimensionality reduction techniques can be classified into two

major families according to the type of procedure that is followed: feature extraction and feature selection. On one hand, *feature extraction* transform the original space of features to an alternative more compact set, while retaining as much information as possible. One shortcoming of this method, that **makes the procedure unsuitable for the purposes of this thesis**, is in the loss of interpretability due to the transformation process that converts the data to a new feature space where a correspondence to the original space cannot be retrieved. By contrast, *feature selection* takes an alternative approach by choosing the best characteristics within the original feature space, rather than relying on transformations of the data [289].

Formally the feature selection problem can be described as follows: Let $\mathcal{X} = \{\vec{x}_1, \dots, \vec{x}_n, \dots, \vec{x}_N\}$ be a set of observations, where each $\vec{x}_n$ is a vector of features extracted from a certain frame n and each $\vec{x}_n$ is associated to its corresponding label ℓ_n from a vector of labels $\vec{\ell} = \{\ell_1, \dots, \ell_n, \dots, \ell_N\}$. It is possible to create a matrix of features $\mathcal{N}_{N \times d}$ by simply concatenating the observations, so that the i -th row is associated to the corresponding i -th element in $\vec{\ell}$. If this matrix is rewritten in terms of its columns instead of its rows, then $\mathcal{N}_{N \times d} = (\vec{\eta}^1, \dots, \vec{\eta}^i, \dots, \vec{\eta}^d)$, where $\vec{\eta}^i$ represents a particular characteristic and d is the dimensionality of the space. The goal of the selection algorithms is to choose an optimised subset $\hat{\mathcal{N}}_{N \times \hat{d}}$ such that $\hat{d} < d$, without degrading the performance of AVQA systems. The methods that are used to accomplish this task can be divided into three major categories [290]: (i) *Wrapper methods*, characterised by using the performance of a particular classifier or regressor for selecting the optimised space. This provides significant advantages in generalisation, at the expense of a considerable computational cost and a certain bias to the classifier that is used. (ii) *Embedded methods*, that exploit the structure of the data to guide the selection process. These methods are not as computationally expensive as the wrapper and are less prone to overfitting, but use strict model structure assumptions. The latter property **makes them undesirable for the purposes of this thesis and therefore are not considered any further** for the sake of simplifying analyses. (iii) *Filter methods* evaluate statistics extracted from the data, that are independent of any particular classifier. As a result, the selected features are generalist. They also have a reduced computational burden compared to both embedded and wrapper techniques. Two main categories exist within the filter methods, one that ranks variables through correlation coefficients and other that considers subset selection [291].

For the purposes of dimensionality reduction, this thesis employs three filter selection algorithms. Notwithstanding, some methods derived of the filter variable ranking, through correlation coefficients, are also employed for other analyses different to selection of features. All these techniques are introduced in the next subsections.

8.3.1 Variable ranking through correlation coefficients

Variable ranking through correlation coefficients methods are not necessarily used to build predictors, but they are often employed to discover

relationships, or as principal or auxiliary selection mechanisms in feature selection because of their simplicity, scalability, and good empirical success [291]. The disadvantage of these methods is the inherent inability to measure relationships between sets of features, as they only quantify association between pairs of variables. Correlation coefficients measure the strength of the association between variables by means of correlation indexes, which often range from 0 to 1 in absolute value, and where the larger the size the stronger the relationship [292]. The most common technique for measuring correlation is the *Pearson correlation* (ρ). This parameter can be viewed as a normalized measure of the covariance of two random variables $\mathcal{A}$ and $\mathcal{B}$ such that:

$$\rho(\mathcal{A}, \mathcal{B}) = \frac{\text{cov}(\mathcal{A}, \mathcal{B})}{\text{std}(\mathcal{A}) \text{std}(\mathcal{B})} \quad (8.19)$$

where $\text{cov}(\cdot, \cdot)$ stands for the covariance operator and $\text{std}(\cdot)$ for the standard deviation operator. ρ varies in the range $\rho \leq 0 \leq 1$ and is an indicator of the linear interdependence between variables. An absolute value close to 1 indicates total interdependence on which one variable might be expressed as a linear combination of the other, whereas a value close to 0 indicates independence (only in a linear sense). As shortcomings, ρ should be normally distributed for its calculation to make sense. Furthermore, only linear dependences are measured with this statistic for which its utility is limited.

Other common techniques usually employed to measure the dependency between variables is the *Kendall rank correlation* and the *Spearman correlation*. On one hand, Spearman correlation makes no assumption about the distribution of variables, but only monotonic relationship between variables might be measured. On the other hand, the Kendall coefficient provides an average measure of the agreement between pairs of members, with a distribution that has slightly better statistical properties than its counterparts, but restricted to measuring monotonic relationships. Therefore, all the methods mentioned are of restricted utility for being incapable of measuring nonlinear relationships between variables. Several attempts have been made in the last few years to create measures capable of detecting more complex relationships, among which the *maximum correlation coefficient* [293], *local gaussian correlation* [294], the *Heller-Heller-Gorfine Statistic* [295] or the *distance correlation* ($\rho_{\mathcal{R}}$) [296] are some notable examples. Despite each one of these presents advantages on its own, this thesis uses the distance correlation due to its ability to detect nonlinear relationships, to compare multivariate/univariate variables and due to its simplicity in implementation.

The distance correlation, $\rho_{\mathcal{R}}$, is analogous to a product-moment correlation, but is based on functions of distances between sample elements rather than sample moments as in the case of ρ [296–298]. Having two multidimensional random variables $\mathcal{A}$ and $\mathcal{B}$, such that $\mathcal{A} = \{a_i\}_{i=1}^Q$, $a_i \in \mathbb{R}^{r_a}$ and $\mathcal{B} = \{b_i\}_{i=1}^Q$, $b_i \in \mathbb{R}^{r_b}$; then $\rho_{\mathcal{R}}$ is characterised by the following properties:

- $\rho_{\mathcal{R}}(\mathcal{A}, \mathcal{B})$ is defined for random vectors $\mathcal{A}, \mathcal{B}$ of arbitrary dimensions (r_a not necessarily equal to r_b).
- $\rho_{\mathcal{R}}(\mathcal{A}, \mathcal{B}) = 0$ characterises independence of $\mathcal{A}$ and $\mathcal{B}$.

- $0 \leq \rho_{\mathcal{R}}(\mathcal{A}, \mathcal{B}) \leq 1$
- It does not require assumptions about the data distribution.

Let also $\Xi_{(Q \times Q)} = [e_{ij}]$ be a *pairwise-distance matrix* obtained from calculating the distances of every element of $\mathcal{A}$ and $\mathcal{B}$. $\rho_{\mathcal{R}}$ can be defined as the square root of the following expression:

$$\rho_{\mathcal{R}}^2(\mathcal{A}, \mathcal{B}) = \begin{cases} \frac{\mathcal{V}^2(\mathcal{A}, \mathcal{B})}{\sqrt{\mathcal{V}^2(\mathcal{A})\mathcal{V}^2(\mathcal{B})}}, & \mathcal{V}^2(\mathcal{A})\mathcal{V}^2(\mathcal{B}) > 0 \\ 0, & \mathcal{V}^2(\mathcal{A})\mathcal{V}^2(\mathcal{B}) = 0 \end{cases} \quad (8.20)$$

where $\mathcal{V}^2(\cdot, \cdot)$ is a *distance covariance* and $\mathcal{V}^2(\cdot)$ are *distance variances* of U -centred distance matrices $\hat{\Xi}_{(Q \times Q)}$ defined as:

$$\mathcal{V}^2(\mathcal{A}, \mathcal{B}) = \frac{1}{Q(Q-3) \sum_{i \neq j} \hat{\Xi}_{i,j}^{\mathcal{A}} \hat{\Xi}_{i,j}^{\mathcal{B}}}$$

And $\hat{\Xi}$ is an U -centered distance matrix obtained from Ξ as follows:

$$\hat{\Xi} = \begin{cases} e_{i,j} - \frac{1}{Q-2} \sum_{l=1}^Q e_{i,l} - \frac{1}{Q-2} \sum_{k=1}^Q e_{k,j} + \frac{1}{(Q-1)(Q-2)} \sum_{k,l=1}^Q e_{k,l}, & i \neq j \\ 0, & i = j \end{cases}$$

After having computed $\rho_{\mathcal{R}}$, ρ or any other correlation measures, it is possible to gauge the relationship that exists between variables. For instance, a practical application is in measuring how correlated is a certain feature to another, serving then as estimator of relevance.

8.3.2 Filter subset selection

Filter subset methods depend on the optimisation of a certain criterion, $\mathfrak{k}$, referred as a *relevance index* or *scoring criterion* [290]. The most straightforward approach to measure the relevance of a certain feature is probably by means of the computation of the mutual information between a certain feature and its associated vector of labels $\vec{\ell}$. In this manner, the mutual information score for a certain feature $\vec{\eta}_i$ can be written as follows:

$$\mathfrak{k}^{MMI}(\vec{\eta}^i) = \mathcal{I}(\vec{\eta}^i, \vec{\ell}) \quad (8.21)$$

where $\mathcal{I}(\cdot, \cdot)$ is the mutual information, which for two random variables $\mathcal{A}$ with alphabet $\mathfrak{A}$, and $\mathcal{B}$ with alphabet $\mathfrak{B}$, is as follows:

$$\mathcal{I}(\mathcal{A}, \mathcal{B}) = \mathcal{H}(\mathcal{A}) - \mathcal{H}(\mathcal{A}|\mathcal{B})$$

where $\mathcal{H}(\mathcal{A})$ is the Shannon entropy and $\mathcal{H}(\mathcal{A}|\mathcal{B})$ is the conditional entropy:

$$\mathcal{H}(\mathcal{A}|\mathcal{B}) = - \sum_{a \in \mathfrak{A}} p(a) \sum_{b \in \mathfrak{B}} p(a|b) \log p(a|b)$$

The calculation of the above scoring criterion for the whole range of features, and its subsequent ranking according to their relevance index, constitutes the so called *Mutual Information Maximisation* (MIM) technique. An important limitation of the method is the assumption of independence among features, which turns out to be suboptimal when they are correlated. An alternative approach that tries to correct such shortcoming by accounting for the relevance and redundancy of the features is the *Joint Mutual Information* (JMI), which assumes that the score for a particular $\vec{\eta}_i$ is:

$$k_{JMI}(\vec{\eta}^i) = \sum_{\vec{p}_j \in \underline{N}} \mathcal{I}(\vec{\eta}^i, \vec{p}_j, \vec{\ell}) \quad (8.22)$$

where $\vec{p}_j$ are characteristics from a set $\underline{N}$ that contains those features that have already been chosen relevant in past iterations. The heuristic to select the current feature consists on finding a candidate $\vec{\eta}_i$ that complements the existing features and maximises the score.

A similar criterion that maximises relevance and minimises redundancy, is the *Max-Relevance Min-Redundancy* (mRMR), which defines the scoring criterion as:

$$k_{mRMR}(\vec{\eta}^i) = \mathcal{I}(\vec{\eta}^i, \vec{\ell}) - \frac{1}{d} \sum_{\vec{p}_j \in \underline{N}} \mathcal{I}(\vec{\eta}^i, \vec{p}_j) \quad (8.23)$$

At the end, the result of filter selection is a list of features ranked from the most to the least relevant.

8.4 DISCUSSION

The present section has introduced three stages of great importance in the design of AVQA systems: preprocessing, characterisation and dimensionality reduction. In the preprocessing stage, the concepts of short-time analysis and normalisation are introduced. During the characterisation stage, some sets of characteristics that are typically employed in voice pathology applications are described, i.e., perturbation, spectral/cepstral, complexity and based on modulation spectra. A table introducing all the described sets of features is presented in Table 8.1. Finally, the dimensionality reduction stage discussed the methods followed in this thesis for diminishing the dimension of the original space of features. Only feature selection methods are employed, and within this category, only three filter are considered: MIM, mRMR and JMI. Variable ranking algorithms based on correlation analysis are also described, however, they are not used for dimensionality purposes but for gauging the strength of the relationship between features and labels.

Table 8.1. Features described in this thesis for the characterisation stage in AVQA systems.

Type of parametrization	Features
Perturbation	<i>Harmonics-to-Noise Ratio</i> (HNR)
	<i>Normalized Noise Energy</i> (NNE)
	<i>Glottal-to-Noise Excitation Ratio</i> (GNE)
Spectral-Cepstral	<i>Mel-frequency Cepstral Coefficients</i> (MFCC)
	<i>Smoothed Cepstral Peak Prominence</i> (CPPS)
	<i>Low-to-High Frequency Spectral Energy Ratio</i> (LHR)
	<i>Perceptual Linear Prediction coefficients</i> (PLP)
Modulation Spectra	<i>Modulation Spectra Homogeneity</i> (MSH)
	<i>Cumulative Intersection Point</i> (CIL)
	<i>Rate of Points Above Linear Average</i> (RALA)
	<i>Modulation Spectrum Percentile</i> (MSP)
Complexity	<i>Largest Lyapunov Exponent</i> (LLE)
	<i>Correlation Dimension</i> (D_2)
	<i>Hurst Exponent</i> (h^e)
	<i>Detrended Fluctuations Analysis</i> (DFA)
	<i>Recurrence Period Density Entropy</i> (RPDE)
	<i>Approximate Entropy</i> (ApEn)
	<i>Sample Entropy</i> (SampEn)
	<i>Modified Sample Entropy</i> (mSampEn)
	<i>Gaussian Kernel Sample Entropy</i> (GSampEn)
	<i>Fuzzy Entropy</i> (FuzzyEn)
	<i>Permutation Entropy</i> (PE)
	<i>Rényi HMM Entropy</i> (rHMMEn)
	<i>Shannon HMM Entropy</i> (sHMMEn)

MACHINE LEARNING AND EVALUATION OF RESULTS

After having extracted the relevant features from the speech, the next block in the design of AVQA systems is referred to training decision machines capable of detecting/assessing the condition of voices. As illustrated in Figure 5.1, the stages within the block are the following: (i) *Machine learning and decision making* where the extracted characteristics are employed to train models and provide decisions about the membership of the input speech signal. (ii) *Evaluation of results*, where the performance of the system is evaluated by means of assessment metrics.

The machine learning procedure receives different names depending on the type of decision making task that is involved. For the purposes of this thesis, the tasks involve vectors of features which are associated to a certain label, i. e., given a training set of observations $\mathcal{X} = \{\vec{x}_1, \dots, \vec{x}_n, \dots, \vec{x}_N\}$, where each $\vec{x}_n$ is associated to a label $\vec{\ell} = \{\ell_1, \dots, \ell_n, \dots, \ell_N\}$, the aim of the procedure is to learn a mapping from the input set of observations to the labels. This type of task is known as *supervised learning*; in opposition to *unsupervised learning*, which is related to the discovery of structure in the data in the absence of labels.

According to the nature of the labels, a further categorisation can be considered. In this manner, if these are categorical or nominal the learning task is known as *classification* (e. g., pathological or normophonic); if these are continuous the learning task is known as *regression* (e. g., the age of a patient); if on contrary these are ordinal and discrete the learning task receives the name of *ordinal classification* (e. g., the order in each one of the traits in the GRBAS scale). These three tasks are addressed in the next sections, as well as some specific techniques that are employed in this thesis.

9.1 CLASSIFICATION

The GMM are one of the most popular classifiers in speech based applications. Its popularity arises from the modelling capability they offer and the probability framework on which they stand. Derived from the idea of GMM, other classifiers have been proposed to model and compensate for inter and intra-class variability. The current section is devoted to present the GMM and GMM-based classifiers considered in this thesis.

9.1.1 Gaussian Mixture Models

The **GMM** is a type of generative classifier that has provided excellent results in diverse speech-related applications. The idea behind **GMM** is to model a feature vector $\vec{x}_n$, of dimension d , as a finite mixture of G multivariate Gaussian components of the form:

$$p(\vec{x}_n|\Theta) = \sum_{g=1}^G \lambda_g \mathcal{N}(\vec{x}_n; \vec{\mu}_g, \Sigma_g) \quad (9.1)$$

where λ_g are mixture weights $0 \leq \lambda_g \leq 1$, $\sum_{g=1}^G \lambda_g = 1$; $\mathcal{N}(\cdot)$ are Gaussian density functions with d -dimensional means $\vec{\mu}_g$ and $d \times d$ -dimensional covariance matrices Σ_g ; and $\Theta = \{\lambda_g, \vec{\mu}_g, \Sigma_g\}_{g=1}^G$ comprises the aforementioned set of parameters. The objective is to estimate the parameters Θ that best model the input feature vector.

Extending above definition to a dataset $\mathcal{X}^o = \{\vec{x}_1, \dots, \vec{x}_n, \dots, \vec{x}_{N_o}\}$, of N utterances with the superindex o indicating data with membership to the class o , the value of Θ^o is obtained by means of a *Maximum Likelihood* (**ML**) optimization scheme applied to the likelihood function $p(\mathcal{X}^o|\Theta^o)$:

$$\arg \max_{\Theta^o} \{p(\mathcal{X}^o|\Theta^o)\} \quad (9.2)$$

A closed form solution to such maximization is not possible. For this reason, the *Expectation-Maximization* (**EM**) algorithm is employed to iteratively refine the parameters in Θ^o , hence increasing the likelihood of the estimated model for the observed $\mathcal{X}^o$ [299]. The resulting **GMM** model after the **EM** estimation is of the form:

$$p(\mathcal{X}^o|\Theta^o) = \prod_{n=1}^{N_o} p(\vec{x}_n|\Theta^o) \quad (9.3)$$

When a large training set $\mathcal{X}^o$ is supplied, the **ML** procedure provides a good estimation of parameters in Θ^o . However, the quality of the model is compromised when the training set is limited. It is possible to partially circumvent that problem by employing a larger auxiliary database including most of the speech characteristics under study, and which is modeled using a **GMM** procedure. This ancilliary model is referred as **UBM** [299] and serves as a well-trained initialization model, for which specific models are derived using the provided -and scarcer- training data. In other words, the idea is to take advantage of the well trained Θ^o parameters of the **UBM** models, instead of having to learn all the parameters for the class-specific model. As a result, the derived models are expected to provide a better parameter estimation than by developing a **GMM** using the training data directly.

The outcome models are termed **GMM-UBM** and are typically generated through **MAP** adaptation, where unlike **ML**, there is a term accounting for the prior distribution of the parameters $p(\Theta^o)$:

$$\arg \max_{\Theta^o} \{p(\mathcal{X}^o|\Theta^o)p(\Theta^o)\} \quad (9.4)$$

In practice the prior distribution is set to the **UBM** model distribution [300], permitting an estimation of the adapted parameters by means of the **EM** algorithm. To simplify calculations even further -and is typically performed in speech related applications- the adaptation procedure is restricted to the means of the **GMM-UBM** model, setting the covariance matrix and the weights to the values of the **UBM**. In this manner, the **MAP** adapted **GMM-UBM** model is defined as:

$$p(\mathcal{X}^o | \Theta^o) = \prod_{n=1}^N \sum_{g=1}^G \lambda_g^u \mathcal{N}(\vec{x}_n; \hat{\mu}_g^o, \Sigma_g^u) \quad (9.5)$$

where the superindex u is referred to the parameters of the **UBM** model and $\hat{\mu}_g^o$ is the g -th **MAP**-adapted gaussian component of the mean:

$$\begin{aligned} \hat{\mu}_g^o &= \alpha_g \mathcal{F}_g + (1 - \alpha_g) \vec{\mu}_g^u \\ \alpha_g &= \frac{\mathcal{Z}_g}{\mathcal{Z}_g + \beta} \\ \mathcal{Z}_g &= \sum_{n=1}^{N_o} \mathcal{R}(g | \vec{x}_n) \\ \mathcal{F}_g &= \frac{1}{\mathcal{Z}_g} \sum_{n=1}^N \mathcal{R}(g | \vec{x}_n) \vec{x}_n \\ \mathcal{R}(g | \vec{x}_n) &= \frac{\lambda_g p_g(\vec{x}_n)}{\sum_{j=1}^G \lambda_j p_j(\vec{x}_n)} \end{aligned} \quad (9.6)$$

where $\mathcal{F}_g$ and $\mathcal{Z}_g$ are first and zeroth-order sufficient statistics respectively, $\mathcal{R}(g | \vec{x}_n)$ are the responsibilities representing how probable is the g -th component in explaining $\vec{x}_n$; α_g is a data-dependent adaptation coefficient controlling the balance between old and new estimates, and β is a fixed relevance factor which in speaker recognition problems is usually set to 16 [299]. An additional simplification in the **GMM** and **GMM-UBM** modelling paradigm is established by using diagonal covariance matrices instead of more complicated covariance configurations. Three reasons for this are the following [299]: (i) the modelling given by a full covariance matrices might be equalled by using a larger order diagonal covariance matrix configuration; (ii) **GMM** models employing diagonal matrices are more computationally efficient; (iii) there is a trend on better performance given by the simpler diagonal matrices compared to full diagonal configurations.

As a matter of example, Figure 9.1 exemplifies the derivation of **GMM-UBM** models by **MAP** adaptation of an **UBM**.

9.1.2 GMM-based factorial models

Several classifiers are derived from the idea of **GMM** and **GMM-UBM**. To describe some of its variations it is first necessary to analyse the **MAP** adapted **GMM-UBM** models in a supervector space, where the term *supervector* is referred to a mapping $\psi : \{A\} \rightarrow \mathfrak{m}$; $A = \{\vec{a}_1, \dots, \vec{a}_i, \dots, \vec{a}_G\}$, $\vec{a}_i \in \mathbb{R}^d$, and

Figure 9.1. GMM-UBM model after the MAP adaptation of the UBM models.

$\vec{m} \in \mathbb{R}^{G \cdot d}$. In practice, supervectors are usually achieved by stacking the mean vectors of GMM models such that [17]:

$$\vec{m}^o = \psi \left(\{ \vec{\mu}_g^o \}_{g=1}^G \right) = [\vec{\mu}_1^o, \dots, \vec{\mu}_G^o] \quad (9.7)$$

The advantage of the supervector space is the possibility of employing a factorial analysis to decompose a supervector model into a *generic all-classes* model plus a *class specific offset*. In this manner, and following the simplifications of speech applications regarding the utilization of mean-only adapted parameters and diagonal covariances, a supervector model $\vec{s}^o$, derived from a GMM-UBM is then decomposed as:

$$\vec{s}^o = \vec{m}^u + \vec{\delta}^o$$

where $\vec{m}^u$ is the UBM mean supervector representing what is common to all classes; $\vec{\delta}^o$ is a class-specific offset representative of class o ; and $\vec{s}^o$ follows $\mathcal{N}(0, I)$. Elaborating further, Equation (9.1.2) can be rewritten in terms of the MAP adapted model:

$$\vec{s}^o = \vec{m}^u + D \vec{z}^o \quad (9.8)$$

where $\vec{z}^o$ is a latent variable of dimensions $Gd \times 1$ following $\mathcal{N}(0, I)$; and $D_{Gd \times Gd}$ is a diagonal covariance matrix satisfying $I = \beta D^T \underline{\Sigma}^{-1} D$, and where $\underline{\Sigma}$ is composed by block diagonal entries of the covariance matrices Σ_g for each of the G components of the UBM [301].

From this point, several approaches are defined to exploit the capabilities of factorial models to decompose the input supervector models into simpler constitutive models. A first variation based on MAP adaptations is the *Eigenvoice Adaptation* (EIG), which assumes that a certain supervector model can be represented in a reduced eigenspace as a linear combination of only a few eigenvectors called *eigenvoices* [302]. The eigenvoice adaptation method assumes that there is a matrix $V_{Gd \times r_v}$ with $r_v \ll Gd$, such that:

$$\vec{s}^o = \vec{m}^u + V \vec{y}^o \quad (9.9)$$

where $\vec{y}^o$ is a latent variable of dimensions $r_v \times 1$, following $\mathcal{N}(0, I)$.

9.1.2.1 Compensated factorial models

Following the supervector notation it is possible to elaborate more complex factorial descriptions that are capable of modelling inter or intra-class components of the input supervector. This permits to deal with some of the undesired between and within-classes variability. In speaker recognition, the variability is introduced due to differences in recording conditions between sessions belonging to a same speaker. In *AVQA* systems, the intraclass variability might be introduced due to differences in the types of studied pathologies, or other extralinguistics such as age or sex.

With this in mind, authors in [302] proposed the *Intersession Variability Compensation (ISC)* as a further development of the *MAP* adaptation which might be applied to compensate for intraclass (intersession) variability. This technique models the variation between different recordings of a certain class, and compensates for this variation during enrolment and testing by excluding the sources of session variation in the generation of class specific models. Thus, the assumption in *ISC* is that a particular recording n of a class o , $\vec{s}_n^o$, is modelled as the summation of a class-specific model $\vec{s}^o$ and a session dependent offset $\vec{\delta}_n^o$, where class-specific and session dependent supervectors are statistically independent and normally distributed [302]:

$$\vec{s}_n^o = \vec{s}^o + \vec{\delta}_n^o$$

Assuming that within class variation is contained in a low dimensional linear subspace of the *GMM* supervector, $r_u \ll Gd$, the session-dependent offset can be expressed as:

$$\vec{\delta}_n^o = U \vec{x}_n^o$$

where $U_{Gd \times r_u}$ is a matrix containing the intra-class variation and $\vec{x}_n^o$ is a latent vector of dimension $r_u \times 1$, following $\mathcal{N}(0, I)$. Furthermore, if *MAP* is considered for the term $\vec{s}^o$ (Equation (9.8)), the *ISC* model is expressed as:

$$\vec{s}_n^o = \vec{m}^u + D \vec{z}^o + U \vec{x}_n^o \quad (9.10)$$

A further improvement of *ISC*, which has been successful in numerous speech related applications is the *Joint Factor Analysis (JFA)*. This technique combines the *MAP* and *EIG* adaptation, within the *ISC* framework. In this manner the assumption in *JFA* is the same as in *ISC*, in which a class and a session-dependent supervector $\vec{s}_n^o$ can be decomposed as the sum of two supervectors, one of which depends on the class, $\vec{s}^o$, and the other on the recording $\vec{\delta}_n^o$. Then, and for representing the class-dependant supervector, both *MAP* and *EIG* criterions are combined so that the class-dependent supervector $\vec{s}^o$ is represented as:

$$\vec{s}^o = \vec{m}^u + D \vec{z}^o + V \vec{y}^o$$

In order to incorporate session effects, a model similar to the one presented in Equation (9.1.2.1) is used, yielding a *JFA* model of the form:

$$\vec{s}_n^o = \vec{m}^u + D \vec{z}^o + V \vec{y}^o + U \vec{x}_n^o \quad (9.11)$$

From Equation (9.11) it can be observed that if $V = 0$ and $U = 0$, then the assumption is the same as in MAP adaptation (Equation (9.8)). On the other, if $D = 0$ and $U = 0$, the assumption is the same as in EIG (Equation 9.9) [303]. In practice the matrices V , U and D are learnt from training data using a ML scheme, where $\vec{x}_n^o$, $\vec{y}^o$, $\vec{z}^o$ are learnt at enrolment time using a MAP adaptation approach.

9.1.2.2 *i*-vectors

An approximation related to JFA which has provided impressive results in speech-related applications is the *identity vector* or *IV* approach. Unlike JFA that decomposes the input model into two separate spaces that account for class and session information, IV defines a single variability space by means of a *total variability matrix*, T . This is motivated by the evidences obtained in speaker recognition applications, on which models that are supposed to contain session information solely are contaminated by class information. With this in mind, the IV decomposition is as follows [304]:

$$\vec{s}_n^o = \vec{m}^u + T\vec{w}_n^o \quad (9.12)$$

where $T_{Gd \times r_t}$ is a rectangular matrix of low rank and w_n^o is a latent vector of dimension $r_t \times 1$, called *i*-vector, following $\mathcal{N}(0, 1)$. The whole IV process might be viewed as mapping of the input matrix of observations (of size $N \times d$) to a single vector of fixed-length, as illustrated in Figure 9.2.

Figure 9.2. IV as a transformation of a series of utterances to a single vector.

Since the IV scheme only models supervectors in the low dimensional total variability space, session compensation techniques are still necessary. This poses some advantages such as a decreased computational load due to the smaller dimension of the *i*-vectors compared to standard supervectors [304]. One of the most employed compensation techniques is the *Probabilistic Linear Discriminant Analysis* (PLDA). In this approach, an *i*-vector $\vec{w}_n^o$, is further decomposed, following a simplified version of PLDA as [305]:

$$\vec{w}_n^o = \overline{\vec{\mu}_w} + \Phi \vec{v}^o + \varepsilon_n \quad (9.13)$$

where $\Phi_{r_t \times r_h}$ is an eigenvoice matrix; $\overline{\vec{\mu}_w}$ is the mean *i*-vector across the whole training set and all the classes; $\vec{v}^o$ is a vector of latent factors of dimension $r_h \times 1$ following $\mathcal{N}(0, I)$; and ε_n contains the unexplained residual

Method	Formula	Dimensions
MAP	$\vec{s}^o = \vec{m}^u + D\vec{z}^o$	$D \rightarrow Gd \times Gd$ $\vec{z}^o \rightarrow Gd \times 1$
EIG	$\vec{s}^o = \vec{m}^u + V\vec{y}^o$	$V \rightarrow Gd \times r_v$ $\vec{y}^o \rightarrow r_v \times 1$
ISC	$\vec{s}_n^o = \vec{m}^u + D\vec{z}^o + U\vec{x}_n^o$	$U \rightarrow Gd \times r_u$ $\vec{x}_n^o \rightarrow r_u \times 1$
JFA	$\vec{s}_n^o = \vec{m}^u + D\vec{z}^o + V\vec{y}^o + U\vec{x}_n^o$	
IV	$\vec{s}_n^o = \vec{m}^u + T\vec{w}_n^o$	$T \rightarrow Gd \times r_t$ $\vec{w}_n^o \rightarrow r_t \times 1$
PLDA	$\vec{w}_n^o = \overline{\vec{\mu}_w^o} + \Phi\vec{v}^o + \varepsilon_n$	$\Phi \rightarrow r_t \times r_h$ $\vec{v}^o \rightarrow r_h \times 1$

Table 9.1. Summary of Factorial models

variability of the model. These parameters are typically estimated by means of an EM scheme as in [306].

Since non-Gaussian behaviour of class and session effects in *i*-vector representations are expected [306], a transformation that ensures normality of the input *i*-vectors before the PLDA modelling is desirable. In this regard, the use of a radial gaussianization process in which a linear whitening is followed by a length normalization of *i*-vectors is proposed in [306]. The whitening and normalization aims at postprocessing *i*-vectors so that they are evenly distributed around a unitary hypersphere. As a result, an almost Gaussian distribution is obtained, which resembles more the distribution assumed by the PLDA model [305, 306]. In particular the whitening centralises and transforms the covariance of an *i*-vector into an identity matrix as follows [307]:

$$\vec{w}_n^{o'} = \mathbf{W}(\vec{w}_n^o - \overline{\vec{w}^o})$$

where $\mathbf{W}$ is the transform computed by means of a Cholesky decomposition of the form $\overline{\Sigma_w^{-1}} = \mathbf{W}\mathbf{W}^T$ and $\overline{\Sigma_w^{-1}}$ is the covariance matrix calculated across some training *i*-vectors. At the end, a length normalization is applied as:

$$\vec{w}_n^{o''} = \frac{\vec{w}_n^{o'}}{\|\vec{w}_n^{o'}\|} \quad (9.14)$$

As a manner to summarise all the factorial models discussed so far, Table 9.1 is introduced.

9.1.3 Support Vector machines

One approach aiming to combine the modelling capabilities of GMM, with a discriminative approach such as the one given by SVM is presented by the GMM-SVM.

A **SVM** is defined as a discriminative binary classifier constructed from sums of a kernel function $\mathcal{K}(\cdot, \cdot)$ such that:

$$f(\vec{x}_n) = \sum_{l=1}^L \xi_l c_l \mathcal{K}(\vec{x}_n, \vec{z}_l) + \delta \quad (9.15)$$

where c_l are ideal outputs taking values -1 or 1 , ξ_l are weights such that $\sum_{l=1}^L \xi_l c_l = 0$, $\xi_l > 0$; δ is a learnt constant; and $\vec{z}_l$ are the L support vectors obtained from a training set by an optimization process. Then, by defining $\vec{s}^{ot}$ and $\vec{s}^{op}$ as supervectors of target and non-target **GMM-UBM** models respectively, a linear sequence kernel of the form is defined:

$$\mathcal{K}(\cdot, \cdot) = \sum_{g=1}^G \left(\sqrt{\lambda_g^u} \Sigma_g^{u(-1/2)} \vec{s}^{ot} \right)^\top \left(\sqrt{\lambda_g^u} \Sigma_g^{u(-1/2)} \vec{s}^{op} \right) \quad (9.16)$$

Which is equivalent to perform a weight and covariance rescaling of the input mean supervectors before the **SVM** training. Therefore the model in Equation (9.15) might be rewritten in the form of a **GMM-SVM** as:

$$f(\vec{x}_n) = \left(\sum_{l=1}^L \xi_l c_l \psi(\vec{z}_l) \right)^\top \psi(\vec{x}_n) + \delta \quad (9.17)$$

where $\psi(\cdot)$ is the supervector operator.

9.1.4 Decision making

Having trained the different systems using any of the approaches presented in previous sections, the next step is to generate probabilistic outputs between an input test utterance and a trained model. These probabilistic outputs are also known as *scores* and are employed for the actual classification. The rationale is that the larger the score the larger the probability of the input utterance of being generated by the tested model. However, the decision about the membership is taken by means of a likelihood decision function, $\Lambda(\cdot)$, testing the input utterance against all sets of models. Associated to the decision function, there is a *threshold* value that is trained to take decisions about the membership of the input vector.

For the simplest case of the **GMM** and **GMM-UBM** models, and having two trained models with parameters Θ^{ot} for a target class o_t , and Θ^{op} for the non-target class o_p ; a derived expression for deciding the membership of an input test data $\vec{x}'$ of unknown class is as follows:

$$\Lambda(\vec{x}') = \log p(\vec{x}' | \Theta^{ot}) - \log p(\vec{x}' | \Theta^{op}) \quad (9.18)$$

For the **ISC** and **JFA** case, a point estimate scoring is discussed in [308], which is as follows:

$$\Lambda(\vec{x}') = SC(\vec{x}', \Theta^{ot}) - SC(\vec{x}', \Theta^{op}) \quad (9.19)$$

and:

$$SC(\vec{x}', \Theta^o) = [\mathbf{V} \vec{y}^o + \mathbf{D} \vec{z}^o] \Sigma^{-1} [\vec{F}(\vec{x}') - \mathbf{N}(\vec{x}') \vec{m}^u - \mathbf{N}(\vec{x}') \mathbf{U} \vec{x}^o]$$

where $\vec{F}(\vec{x}')$ is a $Gd \times 1$ vector obtained after concatenating $\mathcal{F}$ (first order statistics), and $N(\vec{x}')$ is a diagonal matrix composed of blocks of $\mathcal{Z}$ (zero-th order statistics).

Regarding the [IV](#) approach, if the system does not include the session compensation stage, the log-likelihood is simply calculated as the angle between a model i -vector, $\vec{w}^o$, which is the mean i -vector representing the class, and the test i -vector, $\vec{w}'$, as:

$$\Lambda(\vec{x}') = \frac{\langle \vec{w}', \vec{w}^{o_i} \rangle}{\|\vec{w}'\| \|\vec{w}^{o_i}\|} - \frac{\langle \vec{w}', \vec{w}^{o_p} \rangle}{\|\vec{w}'\| \|\vec{w}^{o_p}\|} \quad (9.20)$$

If a [PLDA](#) scheme is considered, the decision score is rewritten as:

$$\Lambda(\vec{x}') = \frac{p(\vec{w}', \vec{w}^{o_i} | H_0)}{p(\vec{w}' | H_1) p(\vec{w}^{o_i} | H_1)} - \frac{p(\vec{w}', \vec{w}^{o_p} | H_0)}{p(\vec{w}' | H_1) p(\vec{w}^{o_p} | H_1)} \quad (9.21)$$

where $p(\vec{w}' | H_0)$ is the probability of both models coming from the same class and $p(\vec{w}' | H_1)$ that they come from different classes.

Finally, for the [GMM-SVM](#) approach, equation (9.17) can be rewritten in such a manner that the decision function is simply a dot product:

$$\Lambda(\vec{x}') = \omega^\top \psi(\vec{x}') + \delta \quad (9.22)$$

where $\omega = \sum_{l=1}^L \xi_l c_l \psi(\vec{z}_l)$

9.1.5 Fusion of scores

The scores of different and heterogeneous systems can be combined through the employment of fusion techniques. Typically this enhances performance as it allows to incorporate complementary information that might help to improve efficiency. The most popular fusion approach is based on logistic regression. In this manner, and for K recognisers $\mathfrak{H}_k$ outputting a sequence of scores, the output of a logistic regressor, $\underline{\mathfrak{H}}$, is of the form [\[309\]](#):

$$\underline{\mathfrak{H}} = \sum_{k=1}^K j_k \mathfrak{H}_k + \vec{\mathfrak{I}}$$

where j_k is a positive scalar and $\vec{\mathfrak{I}}$ is a vector of the same dimension as the output of the recognizers. The fusion procedure consists in finding the coefficient j_k and $\vec{\mathfrak{I}}$ by means of a training database.

9.2 ORDINAL CLASSIFICATION

Ordinal regression or classification is a machine learning problem that aims to predict labels which possess an ordered arrangement. As in the classification problem there exist a features matrix $\mathbf{N}$, associated to a vector of labels $\vec{\ell}$. However, the objective of ordinal regression is to find a classification rule to predict the categories to which an observation belongs, accounting that each label only has a finite set of values, i.e., $\ell_i \in \{C_1, C_2, \dots, C_q, \dots, C_Q\}$ and a natural label ordering exists, i.e., $C_1 \prec C_2, \dots, C_q \prec C_Q$, where $\prec$ is an order relation operator [\[310\]](#). Two of the ordinal classifiers which are employed in this thesis are presented next.

9.2.1 Proportional Odd model

Proportional Odd Model (POM) is member of a family of models, called cumulative link models, on which cumulative probabilities $p(\ell_n \preceq C_q | \vec{x}_n)$ are estimated considering ordering, and which can be related to standard probabilities as follows [310]:

$$p(\ell_n \preceq C_q | \vec{x}_n) = p(\ell_n = C_1 | \vec{x}_n) + \cdots p(\ell_n = C_q | \vec{x}_n)$$

Under the assumptions that for $q = 1$, $p(\ell_n = C_1 | \vec{x}_n) = p(\ell_n \preceq C_1 | \vec{x}_n)$; for $q = Q$, $p(\ell_n \preceq C_Q | \vec{x}_n) = 1$, and for $q = 2, \dots, Q$:

$$p(\ell_n = C_q | \vec{x}_n) = p(\ell_n \preceq C_q | \vec{x}_n) - p(\ell_n \preceq C_{q-1} | \vec{x}_n)$$

POM is based on a linear transformation of the type $f(\vec{x}_n) = \wp^T \vec{x}_n$, but accounting for the ordering of the labels, which is satisfied with the model:

$$g^{-1}\{p(\ell_n \preceq C_q | \vec{x}_n)\} = b_q - \wp^T \vec{x}_n$$

where g^{-1} is a monotonic function referred as the inverse link function and b_q is the threshold for class C_q . The set of thresholds $\mathbf{b} = \{b_1, b_2, \dots, b_{Q-1}\}$ satisfy the constraints $b_1 \leq b_2 \leq \dots \leq b_{Q-1}$, ensuring that $p(\ell_n \preceq C_q | \vec{x}_n)$ increases with q . The standard logistic function is employed so that:

$$g^{-1}\{p(\ell_n \preceq C_q | \vec{x}_n)\} = \log \frac{p(\ell_n \preceq C_q | \vec{x}_n)}{p(\ell_n \succ C_q | \vec{x}_n)} = b_q - \wp^T \vec{x}_n \quad (9.23)$$

9.2.2 Extreme learning machines with ordered partition.

Extreme Learning Machine with Ordered Partitions (ELMOP) rely on the idea of *Extreme Learning Machine* (ELM), which are single-hidden layer feed forward neural networks whose input weights and biases can be randomly assigned and the output weights can be analytically determined by a simple inverse operation [311]. A graphical depiction of an extreme learning machine is presented in Figure 9.3.

Figure 9.3. Example of an ELM network.

Mathematically, an [ELM](#) network with S hidden nodes and activation function $\kappa(\cdot)$ can be written as follows:

$$\sum_{i=1}^S \vec{W}_i \kappa_i(\vec{p}_i \vec{x}_n + b_i) = \mathfrak{T}_i$$

where $\vec{p}_i$ is the weight vector connecting the i -th hidden neuron and the input neurons; $\vec{W}_i$ is the weight vector connecting the i -th hidden neuron and the output neurons; b_i is the threshold for the i -th neuron. This can be compactly written as:

$$\mathcal{M}\mathcal{W} = \mathfrak{T}$$

And the result can be retrieved as simply:

$$\mathcal{W} = \mathcal{M}^\dagger \mathfrak{T}$$

where $\mathcal{M}^\dagger$ is the Moore-Penrose generalized inverse of the hidden-layer output matrix $\mathcal{M}$.

[ELMOP](#) extend the idea of the [ELM](#) but taking into account the ordering of the labels. To do so, each label is coded according to a modified error correcting output code as in before the actual training of the [ELM](#) network [311]. The method is as follows: (i) for a certain label C_q , Q bits are employed for the coding procedure; (ii) the positions from 1 to q are filled with 1 while the remaining are filled with -1 ; (iii) the result is a coded label matrix $\mathfrak{T} = [\mathfrak{T}_1, \mathfrak{T}_2, \dots, \mathfrak{T}_Q]$ which is then employed to train an [ELM](#).

To illustrate the method, the coding procedure for a set of labels $\vec{\ell}$ with $Q = 5$ is shown in Table 9.2.

Table 9.2. Coding matrix following the method proposed in [311]

Label	Coding bits				
	1	2	3	4	5
1	1	-1	-1	-1	-1
2	1	1	-1	-1	-1
3	1	1	1	-1	-1
4	1	1	1	1	-1
5	1	1	1	1	1

9.2.3 Gaussian mixture regression

Gaussian Mixture Regression ([GMR](#)) follows the idea of [GMM](#), but modelling the data with a joint probability function that relates observations, $\mathcal{X}$, and output labels, ℓ . The training phase of the [GMR](#) consists then, of an unsupervised learning procedure, where a [GMM](#) is fitted to the vector $\vec{\ell}_n = [\vec{x}_n, \ell_n]$, resulting from the concatenation of the training samples and their labels

[312, 313]. The parameters of the GMM are adjusted using the EM, and the result is a GMM as in Equation 9.1:

$$p(\vec{\mathcal{E}}_n | \Theta) = \sum_{g=1}^G \lambda_g \mathcal{N}(\vec{\mathcal{E}}_n; \vec{\mu}_g, \Sigma_g)$$

The parameters $\vec{\mu}_g$ and Σ_g , can now be decomposed into input and output components as follows:

$$\vec{\mu}_g = [\vec{\mu}_{g,x}; \vec{\mu}_{g,\ell}]$$

and:

$$\Sigma_g = \begin{pmatrix} \Sigma_{g,x} & \Sigma_{g,x\ell} \\ \Sigma_{g,\ell x} & \Sigma_{g,\ell} \end{pmatrix}$$

where the subindex x or ℓ is referred to the observation $\vec{x}_n$ and to the label ℓ_n respectively. For prediction purposes and assuming an observation $\vec{x}$, the expected label, $\hat{\ell}$, is computed as:

$$\hat{\ell} = \sum_{g=1}^G h(\vec{x}) \left(\vec{\mu}_{g,\ell} + \Sigma_{g,\ell x} \Sigma_{g,x}^{-1} (\vec{x} - \vec{\mu}_{g,x}) \right) \quad (9.24)$$

where:

$$h(\vec{x}) = \frac{\lambda_g \mathcal{N}(\vec{\mathcal{E}}_n; \vec{\mu}_g, \Sigma_g)}{\sum_{j=1}^G \lambda_j \mathcal{N}(\vec{\mathcal{E}}_n; \vec{\mu}_j, \Sigma_j)}$$

9.3 EVALUATION OF RESULTS

This section introduces some metrics which measures the ability that the decision machines possess to distinguish between labels of different types. Moreover, some techniques which are employed to fuse the scores of diverse systems are described.

A common approach followed in machine learning applications to generalise results and provide valid measures about the actual efficiency of the systems consists in the use of cross-validation methodologies. The basis of these is the decomposition of the available dataset into subsets which are used independently for training and testing purposes. On one hand, the *training partitions* are used to estimate a mapping between observations and labels in the machine learning algorithms. On the other, the *testing partitions* are employed to assess the performance of the machines. This is achieved by comparing the labels that are predicted by the machine learning algorithms with the actual labels.

The manner on how the dataset is decomposed is an important matter. Two of the most popular techniques which are employed in literature, and which are also employed in this thesis are the k -folds and leave-one-out cross-validation. *k-folds cross-validation* derives k disjoint sets of size N/k , called folds, where N represents the number of observations. A total of k iterations are then performed, using in each case a different subset for testing purposes, and the remaining $k - 1$ for training algorithms. The measures of

performance are then evaluated as the mean value calculated among iterations. In the limit $k = N$, the *leave-one-out cross-validation* is performed. In this case, only one observation is used for testing and the remaining of the database is employed for training. This same procedure is repeated N times. Having introduced this validation approaches, it is now possible to derive statistics that evaluate the actual performance given by the systems.

9.3.1 Metrics derived from the performance of the system

The simplest approach for evaluating performance consists on extracting metrics that compare the predicted labels given by the decision machines, and the actual classes. In this manner, if the actual label associated to observation is defined as *target*, and the remaining classes are named *non-target*, four possible scenarios are plausible during this comparison:

- If the instance is of the *target* class and the classification decision is of the *target* class, there has been a correctly identified positive instance leading to a *True Positive (TP)*.
- If the instance is of the *target* class and the classification decision is *non-target*, there has been an incorrectly identified positive instance leading to a *False Negative (FN)*.
- If the instance is of the *non-target* class and the classification decision is of the *target* class, there has been an incorrectly identified negative instance leading to a *False Positive (FP)*.
- If the instance is of the *non-target* class and the classification decision is *non-target*, there has been a correctly identified negative instance leading to a *True Negative (TN)*.

It is possible to represent graphically the above-mentioned decisions using a *confusion matrix*. Assuming a two-class decision, a 2×2 confusion matrix is shown in 9.3.

Table 9.3. Depiction of a confusion matrix

		<i>Actual</i>	
		Positive	Negative
<i>Predicted</i>	Positive	TP	FP
	Negative	FN	TN

Some derived metrics of this confusion matrix include *Sensitivity* (SE) also called *recall* or *True positive rate* (TP_r), *Specificity* (SP), *Accuracy* (ACC), *F-score* (f_β) and *False Positive Rate* (FP_r) which are defined as follows:

$$\begin{aligned} SP &= \frac{\sum TN}{\sum(TN + FP)} \\ SE &= \frac{\sum TP}{\sum(TP + FN)} = TP_r \\ FP_r &= \frac{\sum FP}{\sum(TN + FP)} = 1 - SP \\ ACC &= \frac{\sum(TN + TP)}{\sum(TN + FP + TP + FN)} \\ f_\beta &= \frac{2TP}{2TP + FN + FP} \end{aligned}$$

Other metrics derived from the comparison between the predicted label and the actual label and **that account for the ordinal nature** of the labels include the *Mean Absolute Error* (MAE), which is defined as the average deviation of the predicted class from the true class [314, 315]:

$$MAE = \frac{1}{J} \sum_j |\mathcal{O}[\vec{x}_j] - \mathcal{O}^*[\vec{x}_j]| \quad (9.25)$$

where $\mathcal{O}[\cdot]$ is an operator indicating the position of the label in the ordinal rank, i.e., if a certain label ℓ_n can take up values 0,1,2 and the predicted label is 2, its position is 3; and the superindex $*$ is referred to the position of the predicted label. Similarly, the AMAE is defined as the macro averaged version of MAE [316]:

$$AMAE = \frac{1}{J} \sum_j MAE_j \quad (9.26)$$

where MAE_j is the MAE calculated for instances having class j . The advantage of AMAE compared to MAE is its robustness to cope with data imbalance during ordinal classification. This measure is equivalent to the standard MAE when the datasets are perfectly balanced [316].

To deal with situations where there exist continuous labels, it is possible to modify AMAE and MAE. This is achieved by simply replacing the $\mathcal{O}[\cdot]$ operator, for the predicted and the actual labels. In addition, another standard error measure that is usually employed with continuous labels is the RMSE, which is computed as:

$$RMSE = \sqrt{\frac{\sum(\ell_j - \ell_j^*)^2}{J}} \quad (9.27)$$

where ℓ_j is the actual and ℓ_j^* the predicted label.

9.3.2 Performance curves

Another manner to analyse the performance of **classification systems**, is by means of *Receiver Operating Characteristic* (ROC) and *Detection Error Trade-off* (DET) curves. ROC curves, are two-dimensional representations on which

Figure 9.4. Typical DET curve

TP_r is plotted on the y -axis and FP_r is plotted on the x -axis [317]. Interestingly the ROC curves are insensitive to changes in class distribution, hence, if the proportion of positive to negative instances changes in a test set the ROC curves stay the same. A second type of performance graphic is the DET curve on which both axes represent error rates [318]. It is a variant of the ROC curve on which the primary difference is in the use of the *False Negative Rate* (FN_r) in the y -axis as opposed to the TP_r in ROC, and the employment of logarithmic scales. An example of a DET curve is presented in Figure 9.4, where it can be observed that better performance is obtained as the curve gets closer to the point (0,0) (the blue curve is better than the red one).

An additional measure derived from the ROC and DET curves is the AUC, which is a value ranging between 0 and 1, that is obtained after integrating the ROC curve. A number of reasons favour the use of this metric instead of other classical measures such as the ACC, including: a standard error that decreases as both AUC and the number of test samples increase; decision threshold independence; and invariance to a-priori class probabilities [319].

9.3.3 Assessing the reliability of perceptual evaluations

The assessment of voice pathologies relies on perceptual evaluations, which in practice are analysed according to rating scales. Since assessments are subjective by nature, it is often necessary to compute statistics that provide information about the reliability of the ratings offered by the evaluators. A classical method for computing reliability is based on the *classical-test theory*, which assumes that each speaker possess a *true rating* which is defined as the average of a very large number of ratings carried out by raters randomly chosen from the population, in such a manner that [320]:

$$\text{actual rating} = \text{true rating} + \text{error}$$

Given estimates of the variances, the reliability of an individual rating is summarized by the associated reliability coefficient, which is the ratio of variance of true scores to the variance of actual scores as:

$$\frac{\text{true score variance}}{\text{true score variance} + \text{error variance}}$$

A generalisation of such notions is given by the *generalisability theory* (*G-theory*), which stands on *Analysis of Variance* (ANOVA) to quantify the effects that contribute to the difference between true and actual ratings. In this manner, an ANOVA decomposition of the k -th true rating, r_{ijk} ; where the sub-index i is related to the speaker, and j is related to the rater, is as follows:

$$r_{ijk} = \mu + S_i + R_j + SR_{ij} + e_{ijk}$$

where μ is the grand-mean or the average score over the whole population of speakers and raters, S_i is the difference between the true score for speaker i and the grand mean, R_j represents the average rating bias of rater j relative to the average of all rates, SR_{ij} models the interaction between rater and speaker, and e_{ijk} incorporates the remaining sources of variation. Assuming independence of the different terms, the estimations of the variances can be performed by means of a two-way ANOVA model of random effects model, in such a manner that a decomposition of the variance terms is achieved as:

$$\text{var}(r) = \text{var}(S) + \text{var}(R) + \text{var}(SR) + \text{var}(e)$$

Having this term, it is possible to compute a generalizability coefficient, which quantify the consistency among the multiple measurements on a scale from 0 to 1, and which is defined as [321]:

$$\mathcal{U} = \frac{\text{var}(S)}{\text{var}(S) + \text{var}(R)/N + \text{var}(SR)/N + \text{var}(e)/MN} \quad (9.28)$$

where M is the number of ratings given by the same rater, and N the number of raters.

9.4 DISCUSSION

This section has introduced some of the machine learning and evaluation procedures to be employed in this thesis. In this respect, classification, regression and ordinal classification techniques are described. Regarding classification, GMM-based models are employed due to their good performance in speech-related applications. This include the classical GMM, GMM-UBM and factorial models such as JFA, IV and PLDA for compensation purposes. In addition, the GMM-SVM is also discussed. With respect to the ordinal regression, two methods are described. One is a classical linear method called POM and the other is one based on ELM called ELMOP. Likewise, one regressor which relies on the idea of GMM to predict labels in a continuous fashion is also presented: the GMR. Moreover, some metrics that often employed for evaluation of classifier/regression performance are also discussed. Finally, a metric to measure the consistency between ratings given by evaluators has been included.

Part IV

EXPERIMENTAL SETUP

EXPERIMENTS

The experiments that are followed throughout this thesis are presented next. In this respect, this document considers two major groups of explorations. On one hand, *AVQA systems for the detection of voice pathologies* are studied. Particular attention is given to the influence of several variability factors affecting performance, such as the type of acoustic material, the type of features employed for characterisation or the effect of other extralinguistic traits. In addition, techniques explicitly designed to combat variability are tested out. At the end, the objective is to find the approach providing the best performance in pathology detection. On the other hand, the design of *AVQA systems for the assessment of voice pathologies* -based on perceptual evaluations of the *GRBAS* scale- is also examined. Different sets of features are investigated, aiming at discovering the characteristics that improve the performance of an automatic grading system. Since the assessment of voice pathologies is a continuous problem by nature, the use of perceptual scales based on discrete labels -such as the *GRBAS* scale- constitutes a simplification that might lead to inaccuracies in the characterisation of voice quality. With this in mind, a methodology based on regression is proposed, in such a manner that discrete *GRBAS* labels are mapped to a continuous space on which vocal quality might be better represented.

The current chapter contains the following: (i) first, the acoustic material that is employed for experimentation is reviewed; (ii) next, the different tests within the pathology detection experiment are described; (iii) finally, the tests within the assessment experiment are also presented.

10.1 ACOUSTIC MATERIAL

Three databases containing normophonic and pathological recordings are employed for the major sets of experiments: *HUPA*, *SVD* and *GMar*. In addition, four ancillary datasets are also utilised. The *EUR0M* and *PhoneData* corpora are composed of normophonic registers of speakers reading passages and pronouncing words. The other is the *MEEI* partition. The fourth is a dataset designed for speech recognition purposes containing recordings of sentences uttered in Spanish: the *Albayzin* corpus. Finally, two Spanish corpora of normophonic and dysphonic registers of sustained vowels are considered for evaluation purposes in cross-database experiments: *Aplicación de las Tecnologías de la Información y Comunicaciones* (*ATIC*) and *Hospital Doctor*

Negrín (DN). Another evaluation partition is also defined by considering a subset of the SVD database.

Since one of the objectives (to be described thoroughly later) is to model the perceptual capabilities of a human evaluator, is of great importance to highlight that **all corpora have been evaluated by the same speech therapist**, a professional in voice analysis with more than 20 years of experience.

10.1.1 Hospital Universitario Príncipe de Asturias database

The HUPA database has been recorded by *Universidad Politécnica de Madrid*, Madrid, and by the *Príncipe de Asturias* hospital in Alcalá de Henares, Madrid [322]. It contains the sustained phonation of the vowel /a/ of 366 adult Spanish speakers: 169 pathological and 197 normophonic. It has been recorded using the Kay Computerized Speech Lab Analysis station 4300B with a sampling frequency of 50 kHz and 16 bits of resolution. Pathological registers contain a wide variety of organic pathologies including nodules, polyps, oedemas, carcinomas, etc. A perceptual assessment is performed by a speech pathologist using the GRBAS scale. To summarize some of the statistics of the dataset, Figure 10.1 introduces an histogram containing the distribution per age, sex and condition, and per $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ labels.

10.1.2 Saarbrücken Voice Disorders database

The SVD database holds a collection of speech registers from more than 2000 normophonic and pathological German speakers [43, 87]. It comprises recordings of the sustained phonation of vowels /a/, /i/ and /u/ uttered at normal, high and low pitch, as well as with rising-falling pitch. In addition, it incorporates recordings of the sentence *Guten Morgen, wie geht es Ihnen?* (*Good morning, how are you?*). Registers have been recorded using a sampling frequency of 50 kHz and 16 bits of resolution. A subset of the database has been analysed by a speech therapist to remove registers with a low dynamic range or interferences. After this process, 1538 registers of speakers aged between 16 and 69 years are obtained (568 normophonic and 970 pathological). These recordings are also assessed by means of the GRBAS scale. Figure 10.2 depicts the distribution of the dataset according to the speaker's sex, condition and age, and per $\mathcal{G}$, $\mathcal{R}$, $\mathcal{B}$ labels.

10.1.3 Hospital Gregorio Marañón database

The GMar voice disorders database contains recordings of 272 Spanish speakers uttering the vowels /a/, /i/ and /u/. The database has been recorded using the *MediVoz Captura* system [23] with a sampling frequency of 22050 Hz and 16 bits of resolution. A speech therapist has discarded those registers contaminated by noise, and has assessed the remaining according to the GRBAS scale. Since some of speakers had not recorded the three vowels, the final set of registers generated for each one of the phonated voice

Figure 10.1. Histogram summarising some statistics of the HUPA database: (a) distribution of the labels associated to $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$; (b) distribution of the database according to speaker's sex, condition and age. The colour is referred to the sex of the speaker, whereas the lightness (dark or light) is associated to the condition.

material is as follows: (i) for the vowel /a/ there are 202 audio recordings (95 normophonic and 107 pathological); (ii) for the vowel /i/ there are 190 (94 normophonic and 96 pathological); (iii) whereas for vowel /u/ there are 176 (86 normophonic and 90 pathological) registers.

The partition containing the registers of the vowel /a/ is presented in Figure 10.3, illustrating the distribution of speakers according to their sex, condition and age, as well as per $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ labels.

10.1.4 EUROM database

The EUROM corpora comprises recordings of 60 speakers in each of seven European languages: Danish, Dutch, British English, French, German, Norwegian and Swedish [82]. It has been explicitly designed to aid the phonetic comparison of languages, with similar materials and recording protocols. In this manner, each corpus has been recorded at 20 kHz and 16 bit of resolution in an anechoic room and balancing the acoustic content among the different languages. For this thesis purposes only the German corpus is employed, for which a partition within the corpus called *Many Talker set* is used.

Figure 10.2. Histogram summarizing some statistics of the [SVD](#) database: (a) distribution of the labels associated to $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$; (b) distribution of the database according to speaker's sex, condition and age. The colour is referred to the sex of the speaker, whereas the lightness (dark or light) is associated to the condition.

This subset is conformed by 63 speakers which are asked to perform two tasks:

- Read 100 numbers in the range of 0 to 9999 grouped in 5 blocks of 20.
- Read 5 sentences coming from 40 texts.

Unlike the other databases, this does not contain a [GRBAS](#) perceptual assessment of voice quality. The distribution of speakers per sex and age is presented in Figure [10.4](#).

10.1.5 Albayzin database

The Albayzin database is a Spanish speakers corpora designed for speech recognition purposes [323]. It is composed of three sets: a phonetic set, an application set and a Lombard speech set. Despite these three corpora are available, only the phonetic dataset is of interest for this thesis. The phonetic corpus contains 6800 recordings uttered by 204 speakers. In total 204 phonetically balanced phrases have been employed. Registers have been digitised using a sampling frequency of 16 kHz and 16 bits of resolution.

Figure 10.3. Histogram summarising some statistics of the *GMar* database: (a) distribution of the labels associated to $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$; (b) distribution of the database according to speaker's sex, condition and age. The colour is referred to the sex of the speaker, whereas the lightness (dark or light) is associated to the condition.

Figure 10.4. Histogram summarising the distribution of the *EUROM* database according to speaker's sex and age.

10.1.6 *PhoneDat* database

PhoneDat consists of two corpora: *PhoneDat-I* and *PhoneDat-II*. For this thesis purposes only *PhoneDat-I* is employed (from now on referred simply as *PhoneDat*). This corpus contains 200 phonemically balanced artificial German sentences and two readings, namely "*the North wind*" fable and "*a Butterfly story*" [324]. The database contains around 20000 files which have been ut-

tered by 200 German speakers. All registers have been recorded data at a sampling rate of 16 KHz with a 16 bit encoding.

10.1.7 Massachusetts Ear and Eye Infirmary database

The MEEI voice disorders database contains 710 recordings of English speakers, phonating the vowel /a/ and reading the first sentence of the *Rainbow passage* [83]. The database has been recorded at sampling frequencies ranging from 10kHz to 50 kHz. To ensure a balance in the pathologies under study, a subset of the dataset is chosen in [39]. The resulting partition comprises 226 speakers: 173 pathological and 53 normophonic. The registers have been previously edited to remove the beginning and ending of each utterance, and hence omitting the effects of vowel onsets and offsets.

10.1.8 Hospital Doctor Negrín database

The DN database has been recorded in Las Palmas de Gran Canarias, Spain. It is composed of 181 registers, recorded at 22050 Hz, of Spanish speakers phonating the vowel /a/. Only a partition of 130 registers has been used in this thesis, with some of the statistics of this dataset depicted in Figure 10.5. Since no information about the age of the speakers is provided, the graphic contains the distribution according to the sex, condition, and referred to the distribution of $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ labels.

(a) Distribution of the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ scores.

(b) Distribution per sex and condition.

Figure 10.5. Histogram summarizing some statistics of the DN database: (a) distribution of the labels associated to $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$; (b) distribution of the database according to speaker's sex and condition.

Figure 10.6. Histogram presenting the distribution of $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ labels, in the ATIC database.

10.1.9 Aplicación de las Tecnologías de la Información y Comunicaciones database

The ATIC database contains recordings of 79 Spanish speakers (58 dysphonic and 21 normophonic) phonating the vowel /a/ [325]. Pathological voices have been obtained from public and private Otorhinolaryngology services in Málaga, Spain, whereas normophonic speakers are recorded from teachers and students recruited at Málaga University. Registers have been recorded in quiet rooms under controlled conditions, digitized at 44100 Hz and 16 bits of resolution.

The voice samples have been perceptually labelled by three judges using the GRBAS scale, but omitting $\mathcal{A}$ and $\mathcal{S}$. Perceptual evaluation of all the tracks was carried out by the three judges in a session of approximately 50 minutes. To promote consensus among evaluations two actions are performed:

- A written guide defined the procedure for rating the $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$ traits.
- The quantization levels of each trait are augmented by adding decimals between the 0-1-2-3 categories. Specifically, each interval is split into two sub-intervals. The limits of all the resulting quantization levels are explicitly described in the written protocol. After the completion of the perceptual evaluation of all the voice samples, the decimals are eliminated to provide the commonly used 0-1-2-3 score. The raters unanimously agreed to eliminate decimals by rounding to the nearest integer towards zero. Finally, for each sample that did not get the same rating, a re-evaluation process was carried out in which the judges worked together to give a final rating obtained by consensus.

Besides this rating, an independent assessment is made by the speech therapist whom evaluated the other databases in this paper. The distribution of the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ labels obtained in her assessment are in Figure 10.6.

10.2 EXPERIMENTS

As discussed before, two major sets of experiments are to be examined, one related to voice pathology detection and the other to voice pathology assessment. In order to facilitate comparison of results **the next procedures are always considered before the actual test is performed:**

- To train systems using [UBM](#) models, registers are downsampled to the frequency of the ancillary databases with the lowest sampling rate. In this regard and when considering sustained phonation, registers are down-sampled to 20 kHz, whereas in trials involving running speech, the sampling frequency is set to 16 kHz.
- A *max-normalisation* is applied in such a manner that for a signal $s[t]$; its normalised version, $s_{norm}[t]$, is as follows:

$$s_{norm}[t] = \frac{s[t]}{\max\{s[t]\}}$$

where $\max\{\cdot\}$ represents the maximum value in the register.

Each major experiment (detection or assessment) is composed of several sub-tests that are carried out after considering different experimental conditions. At the same time, each sub-test is composed of trials. In other words, there exists a hierarchy where the experiment is first defined, then the sub-test and lastly the trial. For the sake of clarity, each individual trial is to be referred by a name that describes the type of analysis that is carried out. In this manner, the prefix *D* or *A* is employed to distinguish between the detection or assessment *experiments* respectively. This letter is followed by a number referred to the *sub-test* under analysis. Finally, the particularities of the *trial* are explained within the sub-test by using the symbol "_" followed by further information about the test. As a matter of example, a trial with the name "D1_HUPA" is referred to a detection experiment, under the conditions of the sub-test 1. The particularity of the trial (employment of the [HUPA](#) database) is expressed by the "HUPA" suffix. With these considerations in mind, the remaining of this chapter is devoted to describing all the trials that are examined in this thesis for both detection and assessment tasks.

10.2.1 Detection experiments

Within this category are enclosed all the experiments referred to the classification of dysphonic vs. normophonic voices. The aim is to explore the influence of different variability factors that typically affect the performance of automatic voice pathology detection systems. Additionally, the influence that classification techniques -of great use in the speech recognition field- and which aim to compensate for variability factors is tested out. In total, 4 detection sub-tests are defined as presented in the next paragraphs.

10.2.1.1 Sub-test D1: effects of extralinguistic traits.

The effects of extralinguistic traits (e. g., sex, age) in the design of [AVQA](#) systems is explored. The interest of such an approach is related to the construction of hierarchical-like schemes that address the variability introduced by other extralinguistic factors separately. As a result, the complexity of the voice pathology detection task is decreased, simpler signal processing models are generated, and a subsequent performance improvement is expected.

Figure 10.7 depicts a multi-stage system on which the detection problem is handled separately for each one of the sexes. Then, a further decomposition is carried out to account for the age of the speakers. In this scenario, the extralinguistic information is known beforehand, as it might have been retrieved from the patient’s medical history or by any other means.

Figure 10.7. *Sub-test D1*: hierarchical system based on decomposing the partition of speakers according to sex and age. At the end, several AVQA systems are designed for each one of the subgroups resulting from the decomposition. Note how the male and female children’s partition can be grouped due to the acoustical similarity of both groups.

Four trials which investigate the influence of *sex* and *age* in the detection of voice pathologies are performed. To restrict the number of analyses, only the datasets containing recordings of the vowel /a/ are considered. For analysing the impact of the sex of the speakers in detection tasks, a *sex-independent* and a *sex-dependent* system (an AVQA system for female speakers and another for males) are designed. Similarly, and for studying the influence of the age of the speakers, an *age-independent* and an *age-dependent* detection system are considered. It is noteworthy the significant imbalance in the age distribution of the speakers for all databases as observed from Figures 10.3, 10.1 and 10.2. For this reason, only the HUPA is to be utilised as it holds the most balanced partition of speakers in terms of age. One major difficulty that arises in the design of the age-dependent system comes from having to set limits between age groups (e.g, children are those younger than 16, etc.). Since the major events in vocal quality are related to hormonal and structural changes due to the ageing process, and these vary from person to person, it is certainly impossible to set a unique limit that separates young from older populations. For this reason, and following pragmatic reasons respecting the number of available speakers, the HUPA dataset is divided into two age groups: *older adults* if aged over 50, and *young adults* if younger than 50 and older than 25. The 50 years limit is defined as to include the largest amount of population that have experienced events such as the menopause in females, or the laryngeal ossification and calcification processes in males.

The 25 years inferior limit is set to avoid the impact of hormonal changes due to puberty. A depiction of the age distribution using the proposed partition is presented in Figure 10.8. As observed there is still some imbalance, that is partially reduced when the aforementioned age groups are considered.

Figure 10.8. *Trial D1_HUPA_age*: age distribution for the HUPA partition for patients aged over 25, using bins of 5 years.

The trials which compose the current sub-test are the following:

- *Trial D1_HUPA_sex*: sex-dependent and sex-independent systems trained using the HUPA database.
- *Trial D1_SVD_sex*: sex-dependent and sex-independent systems trained using the SVD database.
- *Trial D1_GMar_sex*: sex-dependent and sex-independent systems trained using the GMar database.
- *Trial D1_HUPA_age*: age-dependent and age-independent systems trained using the HUPA database.

The setup followed for all the trials is presented in Figure 10.9, where the partition block is referred to decomposing the database according to age or sex of the speaker, and each one of the stages is described next:

1. *Preprocessing*: 50% overlapped Hamming windows of 40 ms length are employed as the acoustic material is composed of sustained vowels.
2. *Characterisation*: MFCC features are extracted. The number of MFCC coefficients is varied between 10 and 20 with steps of 2.
3. *Classification*: decision making is carried out using a GMM classifier, varying the number of Gaussian components as $\{2^i\} : i \in \mathbb{Z}; 1 \leq i \leq 9$.
4. *Evaluation of results*: a k -folds procedure has been employed setting the number of folds to $k = 11$ for the experiments evaluating the influence of the sex and $k = 20$ to those where age is explored. The latter is to account for the data imbalance and the reduced number of recordings of the database partitioned according to age. AUC is the preferred

measure of performance evaluation. DET curves and the standard performance metrics for a binary classifier (described in Section 9.3) are also considered.

Figure 10.9. *Sub-test D1: hierarchical detection of voice pathologies.*

10.2.1.2 *Sub-test D2: variability due to the acoustic material and the type of features.*

Sub-test D2 contains the most extensive battery of trials among all experiments. The interest on doing so is two-fold. On one hand, the majority of works in literature explore either the use of sustained phonation, continuous speech or voiced segments extracted from speech signals. However, there are few works that directly compare different types of acoustic material in AVQA methodologies. On the other, and since no single feature can offer enough discrimination power to completely characterise the properties of dysphonia and normophonia, it is of great interest to study which sets -and specifically which characteristics- render the best results in discrimination tasks. This might, potentially, permit the design of robust AVQA systems based on complementary features describing distinct properties of speech production. Likewise, the study of these characteristics with differing speech material might provide hindsight about the consistency of the features.

The acoustic material that is going to be employed includes registers of the vowels /a/, /i/ and /u/ phonated at normal pitch, belonging to the HUPA, SVD and GMar corpora; and recordings of the running speech sentence *Guten Morgen, wie geht es Ihnen?* of the SVD database. It is important to remark that using running speech or sustained phonation affects the type of features that can be tested. In particular, some characteristics cannot be extracted from *running speech* utterances, as they rely on conditions that cannot be always met with such acoustic material. One example is encountered when considering noise features that depend on having a periodic glottal excitation, assumption that cannot be fulfilled on unvoiced segments of speech. Another example is found with some nonlinear features that presume stationarity (condition that is lost during phoneme transitions) or complete determinism (assumption not accomplished when considering unvoiced segments of speech). As an alternative to directly analysing running speech, literature often reports procedures based on retrieving voiced segments of speech

through voiced-unvoiced algorithms [147, 190]. This however implies to disregard information that might be considered informative, specially in certain pathologies when articulation is of importance.

In the most generic case, for sustained phonation or voiced segments extracted from speech, 4 major sets of features can be computed as introduced in Table 10.1 and as defined next:

1. The *Perturbation* (**Pert**) set encompasses **NNE**, **CHNR** and **GNE**.
2. The *Spectral/Cepstral* (**SpecCeps**) set contains **PLP**, **MFCC**, **CPPS** and **LHr**.
3. The *Modulation Spectrum* (**MSs**) set is composed of **MSH**, **CIL**, **RALA**, **MSP₂₅**, **MSP₇₅** and **MSP₉₅**.
4. The fourth approach defines subsets for either the *Uniform Embedding* (**U_e**) or *Non-Uniform* (**NU_e**) sets. The subdivision is taken into account to group characteristics measuring the same properties. Hence:
 - a) A first subset comprises *Dynamic Invariants* (**DynInv**) extracted from the attractor such as the classical dimensionality estimators **D₂**, **LLE**, and another invariant characterising recurrence properties of the attractor such as the **RPDE**.
 - b) The second sub-set consists of features extracted directly from the time series and which measure long-range correlations: the *Long-range Correlation* (**LongRange**) set. It includes **h^e**, and **DFA**.
 - c) The third subset includes *Regularity Features* (**Reg**) which are based on entropy-like quantifiers and that might be extracted from both the time series or from the embedded attractor. It encompasses **ApEn**, **SampEn**, **mSampEn**, **GSampEn** and **FuzzyEn**.
 - d) Finally, the fourth subset includes *Entropy Features based on the Reconstruction* (**EntAtt**) such as **PE**, **rHMMEn** and **sHMMEn**.

Table 10.1. Sets and subsets of features employed during this thesis.

Set	Subset	Features					
Pert	-	NNE	CHNR	GNE			
SpecCeps	-	PLP	MFCC	CPPS	LHr		
MSs	-	MSH	CIL	RALA	MSP ₂₅	MSP ₇₅	MSP ₉₅
U_e/NU_e	DynInv	D_2	LLE	RPDE			
	LongRange	h^e	DFA				
	Reg	ApEn	SampEn	mSampEn	GSampEn	FuzzyEn	
	EntAtt	rHMMEn	rHMMEn	FuzzyEn			

By contrast, running speech can be only characterised with the **SpecCeps** set described above (when no voiced/unvoiced detector is employed). Having this into account, the trials within the current sub-test are described next:

1. *Sub-test considering sustained phonation*: these trials are based on the analysis of the sustained vowels /a/, /i/ and /u/, and employing the totality of features presented in Table 10.1. The trials are as follows:

- *Trial D2_HUPA-A*: considers registers of the vowel /a/ belonging to the [HUPA](#) partition.
 - *Trial D2_SVD-A*: considers registers of the vowel /a/ belonging to the [SVD](#) partition.
 - *Trial D2_GMar-A*: considers registers of the vowel /a/ belonging to the [GMar](#) partition.
 - *Trial D2_SVD-I*: considers registers of the vowel /i/ belonging to the [SVD](#) partition.
 - *Trial D2_GMar-I*: considers registers of the vowel /i/ belonging to the [GMar](#) partition.
 - *Trial D2_SVD-U*: considers registers of the vowel /u/ belonging to the [SVD](#) partition.
 - *Trial D2_GMar-U*: considers registers of the vowel /u/ belonging to the [GMar](#) partition.
2. *Sub-test considering running speech*: These trials are based on the analysis of the running speech recordings of the [SVD](#) database:
- *Trial D2_SVD-RS*: Trial using the sentence of the [SVD](#) corpus and the [SpecCeps](#) set.
 - *Trial D2_SVD-RS-Vd*: Trial using the voiced segments of the sentence and the totality of features presented in Table 10.1. For the detection of voiced segments, the *Munich Automatic Segmentation system* ([MAUS](#)) software has been employed.

Each trial follows the methodological stages presented in Figure 10.10, which are described next:

1. *Preprocessing*: for the trials based on sustained phonation different window lengths are utilised depending on the type of characteristic under study. In this manner, Hamming windows of 40 ms length are employed for [Pert](#) and [SpecCeps](#) sets as suggested in [85]. Hamming windows of 55 ms length are used in the U_e and NU_e sets as in [201]; whereas for [MSs](#) experiments, Hamming windows of 180 ms length are utilised as in [219, 220]. For the trials involving running speech, two scenarios are considered. First, when retrieving voiced segments, the window length is fixed as in the case of sustained phonation. The second is when using the running speech material directly. In this case, the window length is varied in the range of 10 to 30 ms with 5 ms steps. These values are selected not to violate the stationarity assumptions. In addition, silence removal procedures are employed to avoid the characterisation of unwanted segments which might undermine performance.
2. *Characterisation*: the following degrees of freedom are considered for both running speech or sustained phonation experiments. Within the [SpecCeps](#) set, the number of [PLP](#) and [MFCC](#) coefficients are varied in the range of 10 to 20 with steps of 2 coefficients. Considering the [MSs](#) set, the [MS](#) are calculated using 1024 modulation and 128 acoustic bands

as in [220]. The U_e and NU_e sets differs only on if the reconstruction uses τ or $\vec{\zeta}^{opt}$ respectively. Both sets requires to fix the dimension d_e to allow a proper comparison of results. To make this possible, d_e has been computed in a short-time analysis procedure using 55 ms Hamming windows, using all the available recordings of the /a/ vowels in the HUPA, SVD and GMar databases. A histogram is then computed as shown in Figure 10.11. The maximum value obtained among all registers is $d_e = 8$, and thus, is fixed as the embedding dimension for the remaining of the tests. Regarding the Reg subset, the tolerance r is set to $0.35 \text{ std } \{s[t]\}$ as in [201], where $\text{std}(\cdot)$ stands for the standard deviation. In the same manner, to reconstruct the state vectors for the calculation of the regularity estimators, the delay used in the reconstruction (τ or $\vec{\zeta}^{opt}$) is employed, since this permits the avoidance of autocorrelation effects that might compromise the calculation of the features [279]. By contrast, the dimension d_m can be set either to the same value of d_e (thus making the regularity estimate dependent on the embedding), or simply setting $d_e = 2$ as it is often considered in literature [274]. For the sake of completeness both d_m parameters are to be tested, and thus the results are to be referred to either $U_e-d_m=2$ or $U_e-d_m=d_e$. With respect to the EntAtt subset, the number of states n_θ is fixed to 6 and the number of symbols n_v is fixed to 32 as in [201].

3. *Classification*: GMM classifiers are employed since the objective of the current experiment is not to obtain the best absolute performance but to compare the outputs of different systems. The number of Gaussian components is varied in the set $\{2^i\} : i \in \mathbb{Z}; 1 \leq i \leq 9$.
4. *Evaluation of results*: AUC is the preferred measure of performance evaluation. A k -folds procedure has been employed setting the number of folds to $k = 11$. DET curves and the standard performance metrics for a binary classifier described in section 9.3 are considered for each one of the aforementioned sets of features. In addition ρ and $\rho_{\mathcal{R}}$ are also obtained as to measure the degree of correlation between each characteristic and their class membership (normophonic or dysphonia).

10.2.1.3 Sub-test D3: testing out the performance of GMM-based classifiers

This third sub-test investigates the performance of UBM-based classifiers in conjunction with MFCC features, a setup that has been proven useful in several speech recognition tasks but that has not been widely explored in AVQA systems. It is expected that such classifiers compensate for all the sources of variability affecting the classification task, hence improving performance. In this experiment only the registers of the vowel /a/ and running speech belonging to the SVD partition are examined. The experiment using sustained phonation is named *Trial D3_SusPho* and explores three conditions that are defined in accordance to the auxiliary databases that are used to train the UBM and compensation models:

- *Condition C₁*: employs normophonic registers of the HUPA corpus.

Figure 10.10. *Sub-test D2*: automatic detection of voice pathologies using diverse sets of features and acoustic material.

Figure 10.11. *Sub-test D2*: histogram of the values of d_e using utterances of the sustained vowel /a/ of the SVD, HUPA and GMar databases, using short-time analysis and windows of length 55 ms. The maximum value that is observed is 8, thus, it is set as d_e for the embedding tasks.

- *Condition C₂*: uses normophonic and dysphonic recordings of HUPA, MEEI and GMar (vowel /a/ only) to test out the effect of increasing the number of registers.
- *Condition C₃*: employs normophonic data of HUPA, MEEI, GMar and voiced segments of EUR0M, to compare the influence of using acoustic material coming from different speech tasks (running speech, sustained vowels /a/, /i/ or /u/). With respect to EUR0M, the voiced segments are used to match as much as possible to the acoustic content of the SVD partition. Since there is no access to the phonological transcrip-

tion of the audio files, the voiced segments are extracted automatically by means of the MAUS software [326].

For the trial using running speech, *Trial D3_RS*, diverse ancillary databases are employed during the training of UBM and compensation models:

- *Condition C₁*: employs normophonic and dysphonic recordings of the vowels /a/, /i/, /u/ of the HUPA and GMar databases, plus normophonic voiced segments of the EUROM and Albayzin corpora.
- *Condition C₂*: considers normophonic-only recordings of the vowels /a/, /i/, /u/ of the HUPA and GMar databases, plus normophonic voiced segments of the EUROM and Albayzin corpora.
- *Condition C₃*: characterises the normophonic sentences of MEEI, EUROM and Albayzin.
- *Condition C₄*: characterises the normophonic sentences of EUROM.
- *Condition C₅*: characterises the normophonic sentences of EUROM and PhoneDat.

The experimental setup for the *sub-test D3* is introduced graphically in Figure 10.12, whereas each stage is described as follows:

1. *Preprocessing*: for the sustained phonation experiments, 40 ms Hamming windows with a 50% of overlapping between consecutive frames, are employed to ensure that each frame contains at least one pitch period [85]. For the running speech experiments, the length of the window is defined according to the outcomes of *Trial D2_SVD-RS* where this parameter has been varied in the range 10 : 5 : 30 ms.
2. *Characterisation*: MFCC features are extracted from each one of the frames obtained in the preprocessing stage. The number of MFCC coefficients is varied between 10 and 20 with steps of 2 coefficients.
3. *Decision machines*: GMM, GMM-UBM, IV, PLDA and GMM-SVM classifiers are employed. The number of Gaussian components is varied in such a manner that $\{2^i\} : i \in \mathbb{Z}; 1 \leq i \leq 9$. The MAP-adaptation parameter α_g is set to 15, a typical value in speech recognition applications. For IV, the size of the latent factors are varied as $r_t = [20, 35, 50, 65, 80, 95]$; whereas for PLDA the size of the latent factors are varied in the range $r_h = [10, 25, 40, 55]$.
4. *Evaluation of results*: a k -folds procedure has been employed setting the number of folds to $k = 11$. Due to the data imbalance, ACC might not be the best measure of performance since it might bias results towards the majority class. To avoid such a problem, AUC is the preferred measure of performance evaluation [327]. In addition to AUC, DET curves, and the standard performance metrics for a binary classifier described in section 9.3 are considered. It is worth to remark that only the best results within conditions are to be depicted in the DET curves, as the interest is on getting to know which acoustic material provides the best performance.

Figure 10.12. *Sub-test D3*: automatic detection of voice pathologies using **UBM**-based classifiers. Recordings of the sustained vowel /a/ belonging to the **SVD** database are used as target. r_h and r_t are the sizes of the latent factors in the **IV** and **PLDA** models respectively.

10.2.1.4 Sub-test D4: Combination of the best systems.

The last detection sub-test aims at designing an **AVQA** system built around the lessons learnt with the first three sub-tests. Presumably it is to be of a hierarchical type, if including a-priori information about other extralinguistics traits turn out to be informative according to the results of *sub-test D1*. The best features chosen in *sub-test D2* are also employed. If any of the GMM-based classifiers of *sub-test D3* is also useful in improving performance, then it will also be included. At the end, the idea is to design an **AVQA** system capable of generalising results using the most consistent information obtained in all the previous sub-tests. In addition, all the acoustic material available for a certain database is also used to provide a single decision about the condition of speakers. To this end, each one of the **AVQA** systems designed for each speech task are fused via logistic regression.

Two trials are considered in the current sub-test:

- *Trial D4_SVD*, which provides a single decision about the condition of patients by combining the results of the systems based on the vowels /a/, /i/, /u/ and the running speech task of the **SVD** database.
- *Trial D4_SVD_CrossDatabase*, which is designed to test the capabilities of the system presented in *trial D4_SVD* in a cross-database scenario. In particular, the system is assessed using the **HUPA**, **ATIC**, **DN** and **GMar** corpora.

The methodological stages that are followed in both trials are presented in Figure 10.13 and are described next:

1. *Preprocessing*: for the sustained phonation experiments, 40 ms Hamming windows with a 50% of overlapping between consecutive frames are employed, ensuring that each frame contains at least one pitch period [85]. For the running speech experiments, the length of the window is defined according to the outcomes of *trial D2_SVD-RS*. Finally, a hierarchical system is to be used if any of the extralinguistics is informative, according to the results of *sub-test D1*.
2. *Characterisation*: two systems are considered. The first is designed with characteristics providing the best results in *sub-test D2*. Since it is not possible to directly combine sets of features of different type -as windows sizes differ across sets of characteristics- a reparametrisation is performed, modifying the overlap in such a manner that the same number of frames are produced no matter the characteristic that is employed. This procedure is made to modify the overlapping factor but to conserve the original window length. The resulting characteristics are then concatenated and employed for the analysis with the feature selection techniques presented in section 8.3; namely, *MIM*, *JMI*, *mRMR*. To generalise results across databases, acoustic material and features, a scoring procedure is now performed. In this manner and with the results of the ranking techniques and for a certain database, the scoring procedure rewards the best features with a low score, while penalizing the worst with a large value. These scores are then summed up across databases and feature selection techniques. At the end, the features with the lowest scores are regarded as the most informative and consistent and are employed for further testing. A graphic summarising the scoring procedure is introduced in Figure 10.14. In addition to this system, another is designed with *MFCC* features as it is to be used for the *UBM*-based classifiers. For this system, and if including extralinguistic improves performance according to *sub-test D1*, the number of *MFCC* coefficients is varied between 10 and 20 with steps of 2; otherwise the parameter rendering the best results in *sub-test D2* is used.
3. *Decision machines*: *GMM* classifiers are used to train the system employing the most consistent set of features. The *MFCC* system utilises the classifiers that render the highest efficiency as given by the results of *sub-test D3*. The number of Gaussian components is varied in such a manner that $\{2^i\} : i \in \mathbb{Z}; 1 \leq i \leq 9$. The *MAP*-adaptation parameter α_g is set to 15 as it is typically done in speech recognition applications. For *IV*, the size of the latent factors are varied as $r_t = [20, 35, 50, 65, 80, 95]$; whereas for *PLDA* the size of the latent factors are varied in the range $r_h = [10, 25, 40, 55]$.
4. *Fusion of results*: a logistic regression is employed to fuse the system using the most consistent features and a *GMM* classifier, and the system based on *MFCC* and *UBM*-based classifiers. A further fusion is considered to combine the information coming from diverse speech tasks, and thus, to provide a single decision about the condition of speakers.

5. *Evaluation of results:* a k -folds procedure has been employed setting the number of folds to $k = 11$. To allow the fusion of results it has been ensured that each partition contains the same number of recordings, and each one of the generated folds is similar to each other. [AUC](#), [DET](#) curves, and the standard performance metrics for a binary classifier described in section 9.3 are considered.

Figure 10.13. *Sub-test D4:* AVQA system using the best sets of features found in *sub-test D2*, the best classifiers of *sub-test D3* and the extralinguistic traits of *sub-test D1*. A logistic regression fusion procedure is also considered.

10.2.2 Assessment experiments

Within this category are enclosed all the experiments referred to the assessment of dysphonic voices. The goal is to design AVQA systems capable of estimating the level of dysphonia according to the GRBAS scale (disregarding $\mathcal{A}$ and $\mathcal{P}$ due to their lack of consistency). The assessment experiments are performed on the partitions of the vowel /a/ of HUPA, SVD and GMar datasets. In addition, the ATIC and DN databases are to be used for evaluation purposes. Four sub-tests, which are explained in the following subsections, are considered in this experiment.

10.2.2.1 Sub-test A1: Prediction of $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ using different sets of features.

Sub-test A1 considers the relationship between diverse sets of features and their $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ gradings. Since all databases to be investigated are of sustained phonation, the features that are presented in Table 10.1 can be also employed in here. Three trials are defined:

Figure 10.14. *Sub-test D4*: scoring procedure for selecting the most consistent characteristics among databases, feature selection techniques and acoustic material.

- *Trial A1_HUPA*: considers recordings of the vowel /a/ belonging to the HUPA dataset.
- *Trial A1_SVD*: employs the registers of the sustained phonation /a/ in the SVD database.
- *Trial A1_GMar*: uses recordings of the vowel /a/ belonging to the GMar dataset.

The setup followed for all the trials during *sub-test A1* is presented in Figure 10.15 and can be described as follows:

1. *Preprocessing*: as in the detection experiments, Hamming windows of 40 ms length are used in the study of Pert and SpecCeps, of 55 ms length in the U_e and NU_e sets, whereas windows of 180 ms length are utilised for the MSs set. The overlap is modified as to ensure that the four groups of features produce the same number of frames.
2. *Characterisation*: the same degrees of freedom considered during *sub-test D2* are followed in the current trial. Hence, PLP and MFCC coefficients are varied in the range of 10 to 20 with steps of 2 coefficients. For the features of the MSs set, the MS is calculated using 1024 modulation and 128 acoustic bands. For U_e and NU_e , d_e has been set to 8, the tolerance r in the Reg features is set to $0.35 \text{ std}(s[t])$. In the EntAtt subset, the number of states n_θ is set to 6 and the number of symbols n_v to 32.
3. *Classification*: ordinal classification is carried out using POM and ELMOP. The number of hidden layers in the ELMOP algorithm is varied in the range of 5 to 95 with steps of 15.
4. *Evaluation of results*: multiclass performance metrics, correlation measures and descriptors accounting for the errors in the ordering are calculated.

Figure 10.15. *Sub-test A1: methodology for predicting $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$, using different sets of features.*

10.2.2.2 Sub-test A2: Generalisation of results through feature selection methods.

The objective of *sub-test A2* is to study deeper the results obtained during *sub-test A1*. The idea is to construct a generalist system, trained using data of the three datasets, with the aim of providing a prediction about the level of dysphonia using the $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$ traits. Moreover, since the GRBAS evaluations are provided by the same specialist, *the system is expected to model the perceptual capabilities of this speech therapist*. In addition, and unlike other automatic grading systems found in literature, it also accounts for the continuous nature of the assessment task. In this manner, and despite being trained using discrete GRBAS ratings, the system outputs a continuous value characterizing the grade of pathology perceived in the input voice. One single system is designed using the data of the three databases. This is later evaluated by means of three cross-database trials:

- *Trial A2_SVD*: trial using a subset of the SVD database for assessing results. At least 50 registers labelled 0, 50 labelled 1, and so on are chosen from the whole corpus. In total 242 recordings are selected for evaluation purposes.
- *Trial A2_DN*: trial using the 130 recordings of the DN dataset for assessing results.
- *Trial A2_ATIC*: trial using the 79 registers of the ATIC dataset for assessing results.

The setup used during the trials is presented in Figure 10.16, and details about each one of its parts are presented next:

- *Dimensionality reduction*: three dimensionality reduction filter methods, [MIM](#), [mRMR](#) and [JMI](#), are employed to rank features from the most to the least relevant in the [HUPA](#), [SVD](#) and [GMar](#) datasets. This procedure is repeated for each one of the traits ($\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$).
- *Selection of consistent characteristics*: since the objective is to design a single system capable of generalising results, a scoring procedure is employed to select the best global set of features amongst databases and ranking algorithms. The same scoring procedure used in *sub-test D4*, and illustrated in Figure 10.14 is employed.
- *Selection of most representative segments*: since the reduced matrix of the previous stage contains registers of the three databases, it might be reasonable to select only the most representative utterances describing a certain trait. This process resembles an "outlier removal" procedure which is applied as to find a more compact set of utterances. Due to the multidimensional nature of the features space, traditional outlier removal techniques cannot be applied. Instead, a [GMM](#) classifier is trained with one Gaussian for modelling the multidimensional matrix of features. Then, log-likelihood scores are computed for each utterance (segment) in the supermatrix. Next, scores are analysed by means of percentiles, removing those utterances with scores below the median and conserving the remaining for further testing.
- *Machine learning and evaluation of results*: [POM](#) and [ELMOP](#) ordinal classifiers are utilised, varying the number of hidden layers in the [ELMOP](#) algorithm in the range of 5 to 95 with steps of 15. Multiclass performance metrics, and descriptors accounting for the errors in the ordering are calculated for the ordinal regression classifiers. The system is further analysed by means of a Gaussian regressor that provides continuous evaluations. These are assessed by means of error metrics such as [RMSE](#) and [MAE](#). The number of Gaussians in the regressor is varied in the set $\{2^i\} : i \in \mathbb{Z}; 1 \leq i \leq 5$.

Figure 10.16. *Sub-test A2*: automatic grading system based on the [GRBAS](#) scale, trained using data from [HUPA](#), [SVD](#) and [GMar](#) databases.

10.2.2.3 Sub-test A3: influence of sex in automatic assessment tasks.

Just as in the detection experiments, the current sub-test explores the effects of accounting for the sex of the speaker in the design of automatic assessment systems. In this manner, the same methodology of *sub-test A2* is followed to construct two systems: one trained with the evaluations of female speakers and the other with the evaluation of males. During the testing phase, and depending of the sex of the speaker, one of the models generates a decision about the level of dysphonia of the voice according to the $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$ traits. Then, the results of male and female systems are combined to provide a unique measure indicating the performance of the hierarchical system. Similarly to *sub-test A2* the sex-dependent systems are designed with the *SVD*, *GMar* and *HUPA* corpora, whereas testing is performed in three cross-database trials:

- *Trial A3_SVD*: trial using a subset of the *SVD* database for assessing results. At least 50 registers labelled 0, 50 labelled 1, and so on are chosen from the whole corpus. In total 242 recordings are selected for evaluation purposes.
- *Trial A3_DN*: trial using the 130 recordings of the *DN* dataset for assessing results.
- *Trial A3_ATIC*: trial using the 79 registers of the *ATIC* dataset for assessing results.

The setup used during the trials is the same presented in Figure 10.16 and described during *sub-test A2*.

10.2.2.4 Sub-test A4: clinical validation of Gaussian regression systems.

This experiment is aimed at analysing the reliability of the gradings provided by the Gaussian regressors in *sub-test A2*, as assessed by a speech therapist. Two assessment tasks are defined in this experiment. In the *first task*, the evaluator is asked to provide a *blind* assessment of the registers after having re-imagined the *GRBAS* scale in a continuous fashion. To this end, the registers employed in the previous stage are duplicated and randomised, thus obtaining a total of 484 registers for the *SVD* and 244 recordings for the *DN* dataset. By doing so, it is possible to obtain two assessments per register, and thus analyse the consistency of the specialist. In addition, the *ATIC* database has been evaluated once. An illustration of the Matlab® application presented to the evaluator is pictured in Figure 10.17. This contains a slider which can be moved freely in the range 0 to 3, and that includes labels in values 0, 1, 2 and 3.

During the *second task*, the evaluator is inquired to supply a *qualification* about the accuracy of the labels predicted by the automatic system, during the Gaussian regression procedures of *sub-test A2* (or *sub-test A3* if including information about the sex of the speaker is relevant). If any of the labels is too far from what it can be considered acceptable, the evaluator should move the slider to correct the evaluation. The errors between the labels provided

Figure 10.17. *Sub-test A4*: Matlab[®] application given to the speech therapist to provide a blind assessment about voice quality.

by the evaluator and the achieved by the detector are measured by means of [RMSE](#) and [MAE](#). In this case the application is the same one introduced in Figure 10.17, but having set the values of the sliders to the predicted value of the [AVQA](#) system. This procedure is considered for the [ATIC](#) and the [DN](#) database. In summary, the evaluator is asked to perform the first task once using the [ATIC](#) database, and twice for the [SVD](#) and [DN](#) corpora (by using the randomisation procedure explained previously). Moreover, the second task is performed using the [ATIC](#) and [DN](#) partitions.

10.3 DISCUSSION

The current section has introduced all the trials to be performed in this thesis. Two major sets of experiments are defined: one for the automatic detection, and the other for the automatic assessment of voice pathologies. There is a total of 4 detection and 4 assessment sub-tests, which in turn are composed of several trials. A table summarising all the trials to be performed is presented in Table 10.2. In addition, a summary of the databases and their use is introduced in Table 10.3.

Table 10.2. Summary of all the trials to be performed in this thesis. RS. stands for running speech. In parenthesis the type of speech material that is used.

Experiment	Sub-test	Particularities	Trial	Dataset
Detection	1	Influence of extralinguistics, e. g., age and sex	D1_HUPA_sex	HUPA (/a/)
			D1_SVD_sex	SVD (/a/)
			D1_GMar_sex	GMar (/a/)
			D1_HUPA_age	HUPA (/a/)
	2	Influence of different sets of characteristics and acoustic material	D2_HUPA-A	HUPA (/a/)
			D2_SVD-A	SVD (/a/)
			D2_GMar-A	GMar (/a/)
			D2_SVD-I	SVD (/i/)
			D2_GMar-I	GMar (/i/)
			D2_SVD-U	SVD (/u/)
			D2_GMar-U	GMar (/u/)
			D2_SVD-RS	SVD (RS.)
			D2_SVD-RS-Vd	SVD (voiced segments RS.)
	3	UBM-based classification	D3_SusPho	SVD (/a/ in 3 conditions)
			D3_RS	SVD (RS. in 5 conditions)
	4	Best system	D4	SVD (/a/, /i/, /u/, RS.)
Assessment	1	Analysis of features and subset of features	A1_HUPA	HUPA (/a/)
			A1_SVD	SVD (/a/)
			A1_GMar	GMar (/a/)
	2	Design generalist system	A2_SVD	SVD (/a/)
			A2_DN	DN (/a/)
			A2_ATIC	ATIC (/a/)
	2	Sex-dependent generalist system	A2_SVD	SVD (/a/)
			A2_DN	DN (/a/)
			A2_ATIC	ATIC (/a/)
	4	Clinical validation	A4	SVD, DN and ATIC (/a/)

Table 10.3. Summary of all the databases employed in this thesis.

Database	Speech tasks	Language	Use	GRBAS	Reference
SVD	Vowels /a/ /i/ /u/ at normal pitch; running speech	German	Main experiments	yes	10.1.2
HUPA	Vowel /a/	Spanish	Main experiments	yes	10.1.1
GMar	Vowels /a/ /i/ /u/	Spanish	Main experiments	yes	10.1.3
EUR0M	Running speech	German	Ancillary database	no	10.1.4
PhoneDat	Running speech	German	Ancillary database	no	10.1.6
Albayzin	Running speech	Spanish	Ancillary database	no	10.1.5
MEEI	Running speech	English	Ancillary database	no	10.1.7
ATIC	Vowel /a/	Spanish	Cross-database trials; assessment	yes	10.1.9
DN	Vowel /a/	Spanish	Cross-database trials; assessment	yes	10.1.8
Partition of SVD	Vowel /a/	German	Assessment	yes	10.2.2.2

RESULTS ON VOICE PATHOLOGY DETECTION

This chapter is devoted to evaluate the influence of distinct variability factors on automatic voice pathology detection systems. In total 4 sub-tests are conducted:

- *Sub-test D1*: designed to test the effects of including extralinguistic information, such as age and sex, in hierarchical AVQA systems. The employed methodology has been presented in section 10.2.1.1.
- *Sub-test D2*: designed to test the effects of the speech material in AVQA systems, through a set of experiments involving recordings of diverse nature. Additionally, several sets of characteristics are employed to discover the features that provide the best performance. The methodology that has been followed is presented in section 10.2.1.2.
- *Sub-test D3*: includes the techniques typically employed in speaker recognition for removing variability and its effect on AVQA systems. The SVD database is employed for testing, while different ancillary datasets are utilised for training the auxiliary models. The methodology that has been followed is the one presented in section 10.2.1.3.
- *Sub-test D4*: all the knowledge gained from the previous experiments is employed to design an AVQA system comprising the best set of features encountered in *sub-test D2*, the best classifiers from *sub-test D3*, in a hierarchical scheme as in *sub-test D1*. In addition, scores of the different subsystems employing distinct acoustic material have been also tested out. Finally, a trial in a cross-database scenario is performed to verify the robustness of the proposed system. The methodology that is followed has been presented in section 10.2.1.4.

11.1 SUB-TEST D1: EFFECTS OF EXTRALINGUISTIC TRAITS.

This sub-test follows the methodology presented in section 10.2.1.1. It is devoted to analyse the influence of certain extralinguistic traits in the design of AVQA systems, particularly sex and age. In this respect, performance metrics are calculated for a sex-independent system trained using recordings of both female and male speakers. In the same manner, two independent systems are modelled using female-only and male-only data. Metrics are also calculated for a sex-dependent system that accounts for the combined performance of both female and male models. This same methodology is

followed but considering the effects of age, to design an age-dependent and age-independent system.

11.1.1.1 Trial D1_HUPA_sex

Table 11.1 introduces the performance metrics of a sex-independent and a sex-dependent system trained with registers of the vowel /a/ belonging to the HUPA corpus (see section 10.1.1). Likewise, DET curves for these systems are depicted in Figure 11.1.

Table 11.1. Trial D1_HUPA_sex: performance of the sex-dependent (S.D.) and sex-independent (S.I.) system using the HUPA database.
Fe.: Female; Ma.: Male.

	Subtype	ACC	SP	SE	AUC
S.D.	Fe.+Ma.	73.66 ± 4.47	0.72	0.70	0.81
	Fe.: MFCC(20)	73.45 ± 5.76	0.71	0.75	0.81
	Ma.: MFCC(20)	73.97 ± 7.12	0.69	0.80	0.85
S.I.	MFCC(12)	69.62 ± 4.67	0.71	0.77	0.79

Figure 11.1. Trial D1_HUPA_sex: DET curve using the HUPA database.

In comparison to the sex-independent system ($AUC=0.79$), there is a tendency to a performance enhancement when the sex of the speaker is accounted in the classification ($AUC=0.81$). This is expected since modelling female and male systems separately leads to a larger AUC (0.85 and 0.81 respectively) compared to a sex-independent scenario (0.79).

11.1.1.2 Trial D1_SVD_sex

Results for a sex-independent and a sex-dependent system using registers of the vowel /a/ in the SVD corpus (see section 10.1.2), are presented in Table 11.2, whereas the corresponding DET curves are in Figure 11.2.

The outcomes suggest that, in comparison to a sex-independent system, there is a certain tendency towards an efficiency improvement when the sex is deemed in the design of AVQA systems ($AUC=0.78$ when the sex is considered vs. $AUC=0.77$ when it is not). Unlike previous trial, the female-only models achieve higher detection rates than the ones provided by the male-only models ($AUC=0.79$ vs. $AUC=0.77$).

11.1.1.3 Trial D1_GMar_sex

Table 11.3 contains the performance metrics of a sex-independent and a sex-dependent system using registers of the vowel /a/ in the GMar corpus (see section 10.1.3), while the corresponding DET curves are in Figure 11.3.

Table 11.2. *Trial D1_SVD_sex*: performance of the sex-dependent (S.D.) and sex-independent (S.I.) system using the SVD database.

Fe.: Female; Ma.: Male.

Type	Subtype	ACC	SP	SE	AUC
S.D.	Fe. + Ma.	70.74 ± 2.27	0.68	0.73	0.78
	Fe.: MFCC(18)	72.17 ± 2.91	0.70	0.73	0.79
	Ma.: MFCC(16)	68.68 ± 3.62	0.63	0.71	0.77
S.I.	MFCC(18)	70.48 ± 2.28	0.67	0.73	0.77

Figure 11.2. *Trial D1_SVD_sex*: DET curve using the SVD database.

Table 11.3. *Trial D1_GMar_sex*: performance of the sex-dependent (S.D.) and sex-independent (S.I.) system using the GMar database.

Fe.: Female; Ma.: Male.

Type	Subtype	ACC	SP	SE	AUC
S.D.	Fe.+Ma.	70.79 ± 6.27	0.72	0.70	0.78
	Fe.: MFCC(16)	70.45 ± 7.78	0.69	0.72	0.77
	Ma.: MFCC(10)	71.43 ± 10.58	0.76	0.66	0.82
S.I.	MFCC(20)	69.31 ± 6.36	0.69	0.69	0.77

Figure 11.3. *Trial D1_GMar_sex*: DET curve using the GMar database.

In this case, the same tendency of the previous two trials is observed, i.e., the model considering the sex of the speakers outperform the sex-independent system (AUC=0.78 vs. AUC=0.77). Just as in the case of *trial D1_HUPA_sex*, the results of the male-only models outperform those of the female-only models (AUC=0.82 and AUC=0.77 respectively).

11.1.4 Trial D1_HUPA_age

This is the only trial which accounts for the age of the speaker, as the HUPA database has the best balance in terms of age among all the available corpora. Thus, the dataset is partitioned into two groups: young, if the speaker is younger than 50, and elder if aged over 50 (more details about the reasons of this partitioning have been described in section 10.2.1.1). The performance metrics of an age-independent and an age-dependent system using registers of the vowel /a/ belonging to the HUPA corpus are presented in Table 11.4, while the corresponding DET curves are in Figure 11.4.

The outcomes suggest that when the database is partitioned according to the age of the speaker, there is a certain tendency to better results compared to not using this type of information. The improvement is slight

Table 11.4. *Trial D1_HUPA_age*: performance of the age-dependent (A.D.) and age-independent (A.I.) system using the HUPA database.

El.: Elder adult; *Yg.*: Young adult.

Type	Subtype	ACC	SP	SE	AUC
A.D.	<i>El.+Yg.</i>	72.28 ± 5.37	0.64	0.79	0.81
	<i>El.</i> : MFCC(18)	72.50 ± 9.78	0.55	0.84	0.87
	<i>Yg.</i> : MFCC(20)	72.19 ± 6.42	0.67	0.77	0.81
A.I.	MFCC(12)	70.04 ± 5.49	0.64	0.75	0.80

Figure 11.4. *Trial D1_HUPA_age*: DET curve using the HUPA database.

(AUC=0.81 for the age-dependent vs. AUC=0.80 for the age-independent system), not to mention that there are some zones in the DET curve where the age-independent system provides a better performance than the age-dependent case. Interestingly, the partition of elder adults presented an increased performance compared to the one of younger speakers (AUC=0.87 vs AUC=0.81 respectively).

11.1.5 Discussions on sub-test D1

Despite the simplicity of the trials, some interesting observations have arisen. On one hand, and respecting the inclusion of sex information, the outcomes indicate that accounting for the speaker's sex improves the efficiency in pathology detection tasks. Indeed, in the three datasets there is an absolute performance improvement which varied between 4% (using HUPA) to 0.3% (using SVD) compared to the sex-independent detector. It is worth noting that, in general, the best sex-independent and the best female/-male detection systems establish a distinct number of MFCC coefficients. Ultimately, this may be a symptom of having decomposed the problem according to the speaker's sex that induce to different "optimal" operation points, i. e., the number of coefficients maximising performance for female models is not necessarily the same for male ones or when both sexes are considered in conjunction. This might be explained by the significant differences in the vocal tract of male and female speakers, which in turn has consequences on spectral characteristics of speech (for a discussion about this topic see section 6.1.4). Since MFCC characterise these spectral properties, it is reasonable to find these differing operation points for female and male models. Another interesting observation is that the systems trained with female voices generally performed worse than those trained with male data; occurring in two trials except when the SVD database was considered. This phenomenon has long been recognised in speech-based applications, where a decreased performance is often achieved in systems based on spectral/cepstral characterisation and female speech [328]. Within the AVQA field, authors in [202],

have also reported a reduced performance of systems based on female data. A possible explanation of this phenomenon, might be in the presence of a glottal gap that has been observed for normophonic voices, and which occurs more frequently for female speakers. This glottal gap is often correlated to turbulence, which in turn is employed as a descriptor of additive noise content and breathiness. Notwithstanding, it is difficult to conclude as there is not enough evidence supporting such behaviour and there is even a negative result about this particular.

On the other hand, and with reference to including age information, the outcomes suggest a certain tendency of improved detection results when age is deemed in the design of AVQA systems. However, more testing with properly balanced acoustic material is required, as to provide more understanding about its impact. In any case, the trial serves to provide preliminary results reflecting the positive effect of including age into the design of voice pathology systems. It also serves to expound the difficulties that are encountered when using this extralinguistic trait. In particular -and as previously discussed- there exist open issues on how to define age populations, since there might exist divergences between the chronological and the physiological age of speakers. For instance, structural changes in the vocal tract might arise due to environmental factors such as contamination, profession or smoking, in such a manner that one patient's larynx might resemble that of an older or a younger speaker. Studying how to properly label the age of speakers is certainly important and should be examined more thoroughly.

While there is a certain performance improvement in detection tasks when the sex or age of the speakers is considered, its simultaneous inclusion should provide more hindsight about the influence of these extralinguistic traits as both are closely interrelated. Unfortunately, there is not sufficient acoustic material to perform an analysis of this type due to the large imbalance among groups. Notwithstanding, and just for the purposes of illustrating the effects of imbalance, a short trial is performed, where the SVD corpus (the largest at hand) is partitioned according to sex and age as done previously. The resulting distribution of patients according to sex, age and condition is presented in Figure 11.5a, whereas the performance metrics for such partition are presented in Figure 11.5b. The outcomes indicate that due to this data imbalance there are detection rates that are as high as 97% (for the elder males) but at the cost of low or null SP precisely in the groups with the smallest number of recordings. This imbalance largely impacts the results and thus should be taken into account in any detection experiment.

In general, results evidence that partitioning the database according to a criterion that simplifies the detection task, ultimately improves performance. Outcomes also highlight the necessity of larger and better balanced corpora that fully describe variability factors such as age and sex, as these seem to affect AVQA systems. There is a large room for improving the systems presented in the current sub-test, including a more thorough study of other types of features and classifiers correlating to ageing voices or to female/-male conditions. In addition, other factors influencing performance should also be investigated, including accents, vocal effort, etc. An additional as-

Sex	Age	Condition		Sex	Age	ACC	SP	SE	AUC
		Nor.	Dys.						
Female	Yg.	348	311	Female	Yg.: MFCC(10)	66.77 ± 3.60	0.67	0.67	0.75
	El.	18	232	Female	El.: MFCC(18)	80.80 ± 4.88	0.22	0.85	0.73
Male	Yg.	195	155	Male	Yg.: MFCC(18)	61.43 ± 5.10	0.56	0.68	0.69
	El.	7	272	Male	El.: MFCC(20)	97.49 ± 1.84	0.00	1.00	0.86

(a) Distribution according to sex, age and condition.

(b) Performance metrics.

Figure 11.5. *Sub-test D1*: effects of considering age and sex in an AVQA system using registers of the vowel /a/ belonging to the SVD dataset. (a) Distribution of the speakers according to age, sex and condition, (b) Performance metrics using such partition. *El.*: Elder adult; *Yg.*: Young adult. *Nor.*: Normophonic; *Dys.*: Dysphonic.

pect that might be worth to consider is in the study of methodologies that include more effectively extralinguistic information into the system, be it in the form of an a-priori probability or in a regression-like scheme.

11.2 SUB-TEST D2: VARIABILITY DUE TO THE ACOUSTIC MATERIAL AND THE TYPE OF FEATURES.

This sub-test explores the influence of different speech tasks and diverse sets of features in the performance of an AVQA system. The methodology that is followed has been presented in section 10.2.1.2. A total of 9 trials are performed: 3 using the vowel /a/, 2 with the vowel /i/, 2 with vowel /u/, 1 using the voiced segments of a running speech task, and 1 with spectral/-cepstral techniques applied to a running speech sentence.

11.2.1 Trial D2_HUPA-A

The performance metrics of all the considered sets of features for registers of the vowel /a/ in the HUPA database are presented in Table 11.5. The corresponding DET curves are portrayed in Figure 11.6. For the sake of clarity, only the best result within the Reg subset ($d_e = m$ or $d_e = 2$) is depicted. Moreover, in an attempt to study the degree of relationship between features and their class membership, correlation metrics are introduced in Table A.1.

From the obtained outcomes it can be observed that the best results -in terms of AUC- are provided by the Pert set (0.85), followed by the NU_e -Reg subset using $d_e = 2$ (0.84). Indeed, GNE is the feature having the highest correlation ($\rho_{\mathcal{R}}=0.49$). Notwithstanding, RALA ($\rho_{\mathcal{R}}=0.49$) and in general the whole MSs set presents the largest correlation values with their labels among the trial. Within the complexity features, two dynamic invariants deliver the largest correlation: LLE in the U_e set ($\rho_{\mathcal{R}}=-0.29$) and RPDE in the NU_e set ($\rho_{\mathcal{R}}=-0.16$). Interestingly ApEn correlates better than other regularity measures in spite of being the "simplest" estimator. On contrary, the remaining

Figure 11.6. *Trial D2_HUPA-A*: DET curves for different sets of features using the *HUPA* database and the vowel /a/.

Table 11.5. *Trial D2_HUPA-A*: performance in detection tasks for all the tested sets of features, using the *HUPA* database and the vowel /a/. The best AUC within each set is in gray, while the best absolute value is in red.

Set	Subset	ACC	SP	SP	AUC
<i>Pert</i>	-	76.61 ± 4.30	0.77	0.77	0.85
<i>MSs</i>	-	71.77 ± 4.57	0.74	0.70	0.79
<i>SpecCeps</i>	<i>CPPS+LHr</i>	62.10 ± 4.93	0.60	0.64	0.69
	<i>MFCC</i> (12)	69.62 ± 4.67	0.66	0.74	0.79
	<i>PLP</i> (18)	66.94 ± 4.78	0.58	0.77	0.80
<i>U_e</i>	<i>LongRange</i>	56.18 ± 5.04	0.51	0.62	0.60
	<i>DynInv</i>	65.59 ± 4.83	0.63	0.68	0.75
	<i>Reg</i> ($d_e = 2$)	69.62 ± 4.67	0.65	0.74	0.78
	* <i>Reg</i> ($d_e = m$)	69.89 ± 4.66	0.68	0.72	0.78
	<i>EntAtt</i>	75.00 ± 4.40	0.75	0.75	0.83
<i>NU_e</i>	<i>DynInv</i>	71.24 ± 4.60	0.69	0.74	0.81
	* <i>Reg</i> ($d_e = 2$)	75.00 ± 4.40	0.72	0.79	0.84
	<i>Reg</i> ($d_e = m$)	69.35 ± 4.68	0.59	0.81	0.82
	<i>EntAtt</i>	70.70 ± 4.63	0.69	0.73	0.80

Reg features present a lack of correspondence with their labels as indicated by the extremely low correlation values that range from 0.01 to 0.10. Despite that, some of the largest AUC values in the trial are obtained when using the *NU_e-Reg* subset. Finally, *LongRange* presents some of the worst correlation values, which are in line to a decreased performance compared to the remaining subsets (AUC=0.60). In general, the *NU_e* set outperforms the classical *U_e* features. Regarding the *SpecCeps* characteristics, *MFCC* and *PLP* achieve an almost equivalent performance (AUC of 0.80 and 0.79 respectively). While the correlation of *CPPS* ($\rho_{\mathcal{R}}=0.25$) is in the same order of magnitude of *PLP* and *MFCC*, the efficiency of this feature in combination with *LHr* (AUC=0.69)

is much worse than their counterparts, suggesting that the inclusion of **LHr** is seriously affecting performance.

11.2.2 Trial D2_SVD-A

The performance metrics of all the considered sets of features for registers of the vowel /a/ in the **SVD** corpus are presented in Table 11.6, whereas the corresponding **DET** curves are portrayed in Figure 11.7. Moreover, correlation metrics between features and their labels are introduced in Table A.2.

Table 11.6. *Trial D2_SVD-A*: performance in detection tasks for different sets of features, using the **SVD** database and the vowel /a/ as acoustic material.

Set	Subset	ACC	SP	SE	AUC
Pert	-	68.79 ± 2.32	0.68	0.69	0.78
MSs	-	66.67 ± 2.35	0.60	0.70	0.73
SpecCeps	CPPS+LHr	61.12 ± 2.44	0.61	0.61	0.68
	MFCC (18)	70.48 ± 2.28	0.67	0.73	0.77
	PLP (20)	71.07 ± 2.27	0.70	0.72	0.77
U_e	LongRange	61.31 ± 2.43	0.59	0.63	0.68
	DynInv	63.78 ± 2.40	0.62	0.65	0.73
	Reg ($d_e = 2$)	64.56 ± 2.39	0.58	0.68	0.71
	*Reg ($d_e = m$)	67.75 ± 2.34	0.55	0.75	0.74
	EntAtt	68.08 ± 2.33	0.68	0.68	0.75
	DynInv	64.30 ± 2.39	0.60	0.67	0.73
	*Reg ($d_e = 2$)	67.17 ± 2.35	0.64	0.69	0.74
NU_e	Reg ($d_e = m$)	65.54 ± 2.38	0.60	0.69	0.74
	EntAtt	65.02 ± 2.38	0.62	0.67	0.70

(a) **DET** curve for **Pert**, **MSs** and **SpecCeps** sets.

(b) **DET** curve for **U_e** and **NU_e** sets.

Figure 11.7. *Trial D2_SVD-A*: **DET** Curves for different sets of features using the **SVD** database and the vowel /a/.

As in *trial D2_HUPA-A*, the best efficiency is obtained with the **Pert** set (**AUC**=0.78). This is in line with the correlation values provided by **CHNR**

and **GNE** which are among the largest in the trial ($\rho_{\mathcal{R}}=0.35$ and $\rho_{\mathcal{R}}=0.34$ respectively). Similarly, **RALA** achieves the largest correlation in the **MSs** set ($\rho_{\mathcal{R}}=0.31$), but with the remaining features being in the same orders of **RALA**. Nonetheless, the **MSs** set just delivers acceptable results in terms of **AUC** (0.73). The correlation of **MFCC** and **PLP** features is similar ($\rho_{\mathcal{R}}=0.25$), having achieved as well, **AUC** that are close to each other (0.77 and 0.76 respectively). Regarding the complexity characteristics, **RPDE** offers some of the largest correlations in the U_e and NU_e sets. Contrary to the results of the previous trial, h^e provides a large correlation value ($\rho_{\mathcal{R}}=0.26$), as is the case of **PE** in the U_e set ($\rho_{\mathcal{R}}=0.23$). While the correlations of characteristics in the **Reg** set are particularly low (up to $\rho_{\mathcal{R}}=0.11$), acceptable **AUC** values are achieved for both U_e and NU_e sets (**AUC**=0.74). In general, the performance of NU_e and U_e is alike, but with slight differences in the **EntAtt** subset which favours the use of the uniform technique. The U_e -**Reg** performs well when using $d_e = m$, while NU_e -**Reg** performed almost equivalently when using $d_e = 2$ or $d_e = m$.

11.2.3 Trial D2_GMar-A

Table 11.7 summarises the detection results for the different sets of features using registers of the vowel /a/ belonging to the **GMar** dataset; whereas the corresponding **DET** curves are in Figure 11.8. Likewise, correlation metrics of the tested features and their labels are presented in Table A.3.

Figure 11.8. Trial D2_GMar-A: **DET** curves for the best set of features using the **GMar** database and vowel /a/.

Unlike the two previous trials, the best outcomes are obtained by means of the NU_e -**Reg** subset (**AUC**=0.84) and U_e -**EntAtt** (**AUC**=0.83). Notwithstanding, the performance of the remaining complexity subsets is in the same order of magnitude. In general, NU_e achieves better results than U_e . The **SpecCeps**, **MSs** and **Pert** provided acceptable outcomes, with **AUC** ranging from 0.73 to 0.76. The correlation metrics indicate that the highest level of interrelationship between labels and characteristics is provided by **RPDE** ($\rho_{\mathcal{R}}=0.65$) and **mSampEn** ($\rho_{\mathcal{R}}=0.38$) in the NU_e set; and **RALA** ($\rho_{\mathcal{R}}=0.36$). In the NU_e -**Reg** subset, **PE** achieves the largest correlation ($\rho_{\mathcal{R}}=0.21$) as the remaining features range

Table 11.7. *Trial D2_GMar-A*: performance in detection tasks for different sets of features, using the *GMar* database and the vowel /a/.

Set	Subset	ACC	SP	SE	AUC
Pert	-	65.35 ± 6.56	0.65	0.65	0.77
MSs	-	67.82 ± 6.44	0.65	0.70	0.76
SpecCeps	CPPS+LHr	66.83 ± 6.49	0.67	0.66	0.73
	MFCC (20)	69.31 ± 6.36	0.69	0.69	0.77
	PLP (12)	68.81 ± 6.39	0.67	0.70	0.76
U_e	LongRange	56.18 ± 5.04	0.51	0.62	0.60
	DynInv	65.59 ± 4.83	0.63	0.68	0.75
	Reg ($d_e = 2$)	69.62 ± 4.67	0.65	0.74	0.78
	*Reg ($d_e = m$)	69.89 ± 4.66	0.68	0.72	0.78
	EntAtt	75.00 ± 4.40	0.75	0.75	0.83
NU_e	DynInv	71.24 ± 4.60	0.69	0.74	0.81
	*Reg ($d_e = 2$)	75.00 ± 4.40	0.72	0.79	0.84
	Reg ($d_e = m$)	69.35 ± 4.68	0.59	0.81	0.82
	EntAtt	70.70 ± 4.63	0.69	0.73	0.80

from 0.01 to 0.08. Despite the U_e -EntAtt has one of the largest AUC in the trial, the individual features are not very interrelated (except for PE). This is certainly an example of poorly correlated variables that combined altogether achieve good performances in detection tasks.

11.2.4 *Trial D2_SVD-I*

The detection outcomes of having used recordings of the vowel /i/ in the *SVD* dataset are introduced in Table 11.8, while correlation metrics are presented in Table A.4. Similarly, the corresponding DET curves are portrayed in Figure 11.9.

Figure 11.9. *Trial D2_SVD-I*: DET curves for the best set of features using the *SVD* database and vowel /i/.

Table 11.8. *Trial D2_SVD-I*: performance in detection tasks for different sets of features using the *SVD* database and the vowel /i/.

Set	Subset	ACC	SP	SE	AUC
<i>Pert</i>	-	64.37 $\pm$ 2.39	0.64	0.65	0.72
<i>MSs</i>	-	63.20 $\pm$ 2.41	0.64	0.63	0.70
<i>SpecCeps</i>	CPPS+LHr	57.87 $\pm$ 2.47	0.55	0.59	0.62
	MFCC (16)	68.73 $\pm$ 2.32	0.67	0.69	0.76
	PLP (14)	68.21 $\pm$ 2.33	0.67	0.69	0.75
<i>U_e</i>	LongRange	59.04 $\pm$ 2.46	0.57	0.60	0.61
	DynInv	62.29 $\pm$ 2.42	0.62	0.63	0.68
	Reg ($d_e = 2$)	63.39 $\pm$ 2.41	0.51	0.71	0.67
	*Reg ($d_e = m$)	63.46 $\pm$ 2.41	0.62	0.65	0.69
	EntAtt	61.25 $\pm$ 2.43	0.60	0.62	0.65
<i>NU_e</i>	DynInv	63.78 $\pm$ 2.40	0.58	0.67	0.67
	*Reg ($d_e = 2$)	67.49 $\pm$ 2.34	0.58	0.73	0.73
	Reg ($d_e = m$)	61.44 $\pm$ 2.43	0.49	0.68	0.67
	EntAtt	60.47 $\pm$ 2.44	0.52	0.65	0.65

MFCC and *PLP* features accomplish some of the best results in the trial, having *AUC* of 0.76 and 0.75 respectively. Similarly, these features provide some of the largest correlation values in the trial ($\rho_{\mathcal{R}}=0.26$). Within the complexity sets, *NU_e-Reg- $d_e = 2$* achieved the best outcomes (*AUC*=0.73), despite the extremely low correlation values of the individual features (ranging from $\rho_{\mathcal{R}}=0.04$ to $\rho_{\mathcal{R}}=0.07$). As in the previous trials, *Pert* and *MSs* provided favourable outcomes as indicated by *AUC* values superior to 0.70, and some of the largest interrelationships in the trial, provided by features such as *CHNR*, *GNE* and *RALA* ($\rho_{\mathcal{R}}=0.26$ in the three cases).

11.2.5 *Trial D2_GMar-I*

The detection outcomes after having used the *GMar* dataset and the vowel /i/ are presented in Table 11.9, whereas the corresponding *DET* curves are in Figure 11.10. Likewise, correlation metrics between labels and features are presented in Table A.5.

The highest correlation in the trial is obtained using *RALA* ($\rho_{\mathcal{R}}=0.37$), but with remaining features in the *MSs* set of the same order of magnitude. This is in line with the classification results, where the whole set accomplished the best *AUC* in the trial (0.76). Within the *SpecCeps* set, *PLP* and *MFCC* present the largest correlation values ($\rho_{\mathcal{R}}=0.25$ and 0.20), which is translated into a favourable performance in terms of *AUC* (0.65 in both cases). The *Pert* set accomplishes an *AUC*=0.73, having some of the largest correlation values with features such as *CHNR* and *GNE* ($\rho_{\mathcal{R}}=0.28$ in both cases). Respecting the complexity features, while the *U_e-Reg* set provides one of the highest *AUC* (0.73), the correlation values of each one of the features do not exceed $\rho_{\mathcal{R}}$ of

Table 11.9. *Trial D2_GMar-I*: performance in detection tasks for different sets of features using the *GMar* database and the vowel /i/.

Set	Subset	ACC	SP	SE	AUC
Pert	-	66.32 ± 6.72	0.65	0.68	0.73
MSs	-	67.82 ± 6.44	0.65	0.70	0.76
SpecCeps	CPPS+LHr	60.53 ± 6.95	0.62	0.59	0.67
	MFCC (10)	60.53 ± 6.95	0.61	0.60	0.65
	PLP (10)	58.42 ± 7.01	0.60	0.57	0.65
U_e	LongRange	61.05 ± 6.93	0.59	0.64	0.65
	DynInv	64.74 ± 6.79	0.63	0.67	0.70
	Reg ($d_e = 2$)	59.47 ± 6.98	0.48	0.71	0.65
	*Reg ($d_e = m$)	61.58 ± 6.92	0.48	0.75	0.73
	EntAtt	60.53 ± 6.95	0.61	0.60	0.68
NU_e	DynInv	55.26 ± 7.07	0.53	0.57	0.64
	*Reg ($d_e = 2$)	60.00 ± 6.97	0.60	0.60	0.68
	Reg ($d_e = m$)	63.16 ± 6.86	0.64	0.62	0.68
	EntAtt	54.74 ± 7.08	0.54	0.55	0.60

(a) DET curve for *Pert* and *SpecCeps* characteristics.

(b) DET curve for complexity characteristics.

Figure 11.10. *Trial D2_GMar-I*: DET curves for the best set of features using the *GMar* database and vowel /i/.

0.11. In this particular case, the performance given by the U_e set is clearly superior than the one provided by NU_e .

11.2.6 *Trial D2_SVD-U*

Table 11.10 summarises the detection outcomes for the *SVD* dataset and vowel /u/, whereas the corresponding DET curves are portrayed in Figure 11.11. Likewise, correlation metrics between characteristics and labels are presented in Table A.6.

The *SpecCeps* set presents the best results with features such as *MFCC* and *PLP* having an *AUC* of 0.71. Likewise, these two features present some of

Table 11.10. *Trial D2_SVD-U*: performance in detection tasks for different sets of features using the *SVD* database and the vowel /u/.

Set	Subset	ACC	SP	SE	AUC
Pert	-	59.88 $\pm$ 2.45	0.59	0.60	0.66
MSs	-	61.05 $\pm$ 2.44	0.53	0.66	0.67
SpecCeps	CPPS+LHr	58.13 $\pm$ 2.47	0.55	0.60	0.63
	MFCC (16)	65.41 $\pm$ 2.38	0.64	0.66	0.71
	PLP (16)	64.69 $\pm$ 2.39	0.62	0.66	0.71
U_e	LongRange	58.06 $\pm$ 2.47	0.57	0.59	0.64
	DynInv	61.51 $\pm$ 2.43	0.58	0.64	0.68
	Reg ($d_e = 2$)	63.39 $\pm$ 2.41	0.40	0.77	0.65
	*Reg ($d_e = m$)	63.78 $\pm$ 2.40	0.50	0.72	0.67
	EntAtt	59.95 $\pm$ 2.45	0.57	0.62	0.66
NU_e	DynInv	60.86 $\pm$ 2.44	0.58	0.62	0.68
	*Reg ($d_e = 2$)	61.64 $\pm$ 2.43	0.54	0.66	0.66
	Reg ($d_e = m$)	61.44 $\pm$ 2.43	0.55	0.65	0.65
	EntAtt	55.20 $\pm$ 2.49	0.48	0.59	0.59

(a) DET curve for *Pert* and *SpecCeps* characteristics.

(b) DET curve for complexity characteristics.

Figure 11.11. *Trial D2_SVD-U*: DET curves for the best set of features using the *SVD* database and vowel /u/.

the largest correlation values in the trial ($\rho_{\mathcal{R}}=0.19$). Within the complexity sets, and unlike the previous trials, the best outcomes are obtained using the *DynInv* subset ($AUC=0.68$) for both U_e and NU_e . In general, the performance of U_e and NU_e is similar, except for the U_e -EntAtt subset which outperforms its counterpart. As in previous trials some of the largest $\rho_{\mathcal{R}}$ are accomplished by *CHNR* (0.14), *GNE* (0.13) and *RALA* (0.14); which in turn results in *AUC* values in the *Pert* set of 0.66, and of 0.67 in *MSs*.

Table 11.11. *Trial D2_GMar-U*: performance in detection tasks for different sets of features using the *GMar* database and the vowel /u/.

Set	Subset	ACC	SP	SE	AUC
<i>Pert</i>	-	61.36 ± 7.19	0.64	0.59	0.70
<i>MSs</i>	-	63.64 ± 7.11	0.63	0.64	0.72
<i>SpecCeps</i>	CPPS+LHr	59.09 ± 7.26	0.58	0.60	0.68
	MFCC (18)	62.50 ± 7.15	0.62	0.63	0.68
	PLP (16)	67.05 ± 6.94	0.70	0.64	0.73
	LongRange	61.36 ± 7.19	0.64	0.59	0.74
U_e	DynInv	58.52 ± 7.28	0.52	0.64	0.70
	Reg ($d_e = 2$)	55.68 ± 7.34	0.44	0.67	0.65
	*Reg ($d_e = m$)	63.07 ± 7.13	0.60	0.66	0.68
	EntAtt	63.07 ± 7.13	0.67	0.59	0.73
NU_e	DynInv	52.27 ± 7.38	0.49	0.56	0.57
	Reg ($d_e = 2$)	54.55 ± 7.36	0.29	0.79	0.64
	*Reg ($d_e = m$)	57.95 ± 7.29	0.45	0.70	0.65
	EntAtt	52.84 ± 7.38	0.55	0.51	0.59

11.2.7 *Trial D2_GMar-U*

Table 11.11 contains the detection results using the *GMar* dataset and the vowel /u/, whereas the corresponding DET curves are presented in Figure 11.12. Similarly, the correlation analysis is introduced in Table A.6.

Figure 11.12. *Trial D2_GMar-U*: DET curves for the best set of features using the *GMar* database and vowel /u/.

Interestingly, the best results are obtained using the *LongRange* features ($AUC=0.74$), which contravenes the tendency observed so far in which this subset did not provide good efficiency. The U_e -EntAtt set achieves, next to *PLP*, the second best *AUC* (0.73 in both cases). By contrast, the NU_e -Reg subset when considering $d_e = m$ provides a better performance than $d_e = 2$, which also contradicts the tendency obtained in all the previous trials. Its notewor-

thy that in general terms the U_e set outperforms NU_e . Finally, MSs and $Pert$ provides favourable results, which is in line to the tendencies observed in previous trials. In terms of correlation, GNE , PLP and $MFCC$ provide some of the largest interrelationships with their labels in the trial, as indicated by values ranging from $\rho_{\mathcal{R}}=0.19$ to $\rho_{\mathcal{R}}=0.22$.

11.2.8 Trial $D2_SVD-RS$

The current trial characterises registers containing running speech by means of spectral and cepstral features (the $SpecCeps$ set). Experiments are performed by varying the length of the Hamming window in the range 10 to 30 ms with steps of 5 ms. Results for each one of the tested features, using the window length providing the best performance, are presented in Table 11.12. Likewise, the DET curves corresponding to the best-performing operation points are pictured in Figure 11.13.

Table 11.12. Trial $D2_SVD-RS$: best operation points using the sentences in the SVD database, for the spectral/cepstral features.

Features	Length	ACC	SP	SE	AUC
$CPPS+LHr$	25 ms	62.26 ± 2.46	0.60	0.64	0.67
$MFCC$ (18)	20 ms	80.32 ± 2.02	0.74	0.84	0.86
PLP (18)	20 ms	77.90 ± 2.11	0.65	0.85	0.85

Figure 11.13. Trial $D2_SVD-RS$: DET curve of the best results.

Both PLP and $MFCC$ features perform almost equivalently as observed in the DET curves, and as evidenced by the closeness in the AUC values (0.85 for PLP and 0.86 for $MFCC$). When these results are compared to the ones obtained with the acoustic material composed of sustained phonation, there is a clear performance improvement favouring the employment of running speech. Indeed, the vowel providing the best performance in the SVD database achieves an AUC of 0.77 using either PLP or $MFCC$, which contrasts with an AUC of 0.86 and 0.85 when using running speech and PLP and $MFCC$ features, respectively. By contrast, the outcomes of the set $CPPS+LHr$ are in the same order of magnitude as when sustained phonation is utilised.

11.2.9 Trial $D2_SVD-RS-Vd$

The current trial evaluates the performance of a system using only the voiced segments of a sentence belonging to the SVD database. The $MAUS$ software is employed for generating automatic transcriptions and identifying the voiced segments of speech. Despite it would be desirable to employ

the same characterisation stage as in the sustained phonation case, there are some practical reasons that makes difficult this type of analysis. In particular, the window length that is employed in some sets of features is too long to allow the extraction of characteristics. This is expected as running speech is composed of fast transitions between silence, voiced or unvoiced sounds and intervals of sustained phonation are not produced frequently. Indeed, for characterisation with the *MSs* set it is necessary to include voiced segments with a window size of 180 ms. However, there are not vocal segments with such length in the *SVD* database. This behaviour is also encountered when analysing complexity features, where the number of segments whose length is 55 ms is greatly restricted. With this in mind, the current trial employs only those sets which make use of windows of 40 ms, i.e., *Pert* and *SpecCeps* to characterise voiced segments of speech. Results for the tested features are introduced in Table 11.13, whereas the corresponding *DET* curves are in Figure 11.14.

Table 11.13. *Trial D2_SVD-RS-Vd*: best results using voiced segments of the running speech sentence of the *SVD* database.

Features	ACC	SP	SE	AUC
<i>Pert</i>	66.69 ± 2.39	0.67	0.67	0.74
<i>CPPS+LHr</i>	56.15 ± 2.52	0.52	0.58	0.58
<i>MFCC</i> (16)	76.96 ± 2.14	0.70	0.81	0.84
<i>PLP</i> (18)	75.55 ± 2.18	0.67	0.80	0.82

Figure 11.14. *Trial D2_SVD-RS-Vd*: *DET* curve of the best results.

The outcomes indicate that the *MFCC* features provide the best results, closely followed by *PLP*. The performance of the *CPPS+LHr* pair is the worst in the trial. Results also indicate that in comparison to the results of the previous trial (where not only voiced segments are included in the system) there is a decreased performance when voiced segments solely are considered.

11.2.10 Discussions on sub-test *D2*

During the current sub-test several sets of features have been employed for detection tasks. Experimentation has been made on several databases of sustained vowels and running speech. By virtue of the obtained results, we can infer that there is no single feature or set of features that always perform better than the others. Indeed, depending on the type of acoustic material or the particularities of the corpus, a certain set of features outperforms the others. Despite that, there are some sets presenting tendencies of good performance. For instance, *Pert* has proven to be useful for detection purposes as ascertained by the large *AUC* values that varied from 0.66 using the vowel

/u/ and the [SVD](#) database, to 0.85 for the [HUPA](#) dataset (or to 0.86 if running speech is considered). In terms of correlation, [Pert](#) also provided some of the largest values in all trials, specially when considering [CHNR](#) and [GNE](#). In a similar fashion, the [SpecCeps](#) set provides large [AUC](#) values that ranged from 0.71 for the the vowel /u/ and the [SVD](#) database, to 0.8 for the [HUPA](#) dataset. In most of the cases (except for the vowel /i/ of the [GMar](#) database), either [PLP](#) or [MFCC](#) outperformed the results of [CPPS+LHr](#). It is noteworthy that there are some cases where [CPPS](#) provided favourable correlations but [LHr](#) did not. As a result, its combination did not improve performance as reported in literature. By contrast, [MFCC](#) or [PLP](#) generally presented some of the largest correlation values in all the trials. Regarding the [MSs](#) set, there are some interesting observations. In general, the performance of the whole subset remains among the highest compared to the other sets, with an [AUC](#) that varies from 0.67 for the vowel /u/ and the [SVD](#) database, to 0.79 using [HUPA](#). Moreover, there is a favourable correlation of all the features in the set with their labels, specially for [RALA](#), which in many cases provided the largest correlation values among the trials. By contrast, the least correlated characteristics in the [MSs](#) set are [MSH](#) and [CIL](#). Finally, and respecting the complexity features, the results indicate that the subset providing the worst general performance is [LongRange](#) as in almost all cases (but for [GMar](#) and vowel /u/), it is surpassed by the remaining subsets. Regarding the complexity sets, the performance of [Reg](#) and [EntAtt](#) tends to be superior to that of [DynInv](#). Interestingly, and regarding the [Reg](#) features, the outcomes suggest the inefficacy of using the default parameters that the literature usually employs for this type of features, i.e., $\tau = 1$ and $d_e = 2$. Indeed, when considering the [U_e-Reg](#) subset the best results are obtained when $d_e = m$. By contrast the best results for [NU_e-Reg](#) are when $d_e = 2$. As a consequence the lags (be it $\vec{\zeta}$ or τ) between both subsets also vary. Despite there is a difference in the results and a discrepancy on the parameters that should be chosen, the conclusion is that d_e and the lags are certainly important, and that the default parameters might not be the best choice when using this type of features. Not discussed in this thesis but equally important is the tolerance parameter r as it will differ depending on the application and/or the type of biosignal that is considered. Literature has reported the importance of accounting for the parameters d_e , r and/or τ [279, 329, 330], but no through analysis has been made on the simultaneous optimisation of the whole set of parameters, on considering a vector of lags, or on exploring the relationship that exists between them. This indeed, remains an area that is worth to study. Respecting to a direct comparison between [U_e](#) and [NU_e](#), it is difficult to establish which set performs better. In most cases (12 occasions) several of the subsets in [U_e](#) outperform the results of [NU_e](#), specially when the vowels /i/ and /u/ are considered. The contrary only occurs in 5 occasions, while the outcomes are almost indistinguishable in 4 cases. The results of the discrepancies might be related to the non-uniform algorithm itself. Indeed, one of the biggest problems that the technique presents is the reliance on a correct choice of d_e , as the optimisation process heavily depends on this parameter. Additionally, the inherently large non-stationarity

of the speech signals, as well as an optimisation process that might get stuck into local minima might be factors confounding the outcomes. As a result of the above-mentioned facts and due to the increased computational burden of the non-uniform technique, **this approach will not be further considered in this chapter.**

In reference to the acoustic material that is employed, some tendencies are observed. Revisiting the results of the trials using sustained phonation, it can be inferred that the vowel /a/ achieves, in all cases, better results than vowels /i/ and /u/, no matter the set that is analysed. When including running speech into the analysis, it is found that the spectral/cepstral features are in general better when the sentence is employed in comparison to the use of a sustained phonation. In particular for the [Pert](#) set, the vowel /a/ provides a better performance than when using running speech.

In an attempt to compare speech material of different nature with the same set of features, Figure 11.15 presents the best DET curves for all the trials involving the [SVD](#) and [GMar](#) databases. MFCC features are depicted as they are calculated in all tests, including those based on running speech. The outcomes suggest that despite the difficulties of comparing results there is a certain tendency to favour the employment of acoustic material based on running speech. In addition, in both cases the vowel /a/ presents a better performance in comparison to using vowels /i/ or /u/.

Figure 11.15. *Sub-test D2*: DET curves using MFCC features and speech material of different nature belonging to (a) [SVD](#) database; and (b) [GMar](#) database.

It is important to highlight that a fair comparison between speech tasks is difficult, as there are certain differences in the amount of acoustic material that is available when speech and voice are compared. As a matter of example and taking the [SVD](#) database, during *trial D2_SVD-RS* there exist 219.000 frames, whereas 36.691 voiced frames are encountered in *trial D2_SVD-RS-Vd*. By contrast, any of the sustained phonation tasks contain 95.422 frames. It is worth noting that despite the increased number of frames of *trial D2_SVD-RS*, there are utterances which are irrelevant for the detection of laryngeal pathologies such as unvoiced sounds, or other units related

to articulatory movements. Interestingly and despite the reduced number of samples in *trial D2_SVD-RS-Vd*, good performances are attained. In a direct comparison of *trial D2_SVD-RS-Vd* and *trial D2_SVD-RS* better results are obtained when no voiced detector is included. This might be simply a consequence of having used an automatic system for the segmentation of the voice recordings, as some errors might be introduced during this process, specially when pathological voice is considered. In any case, disregarding the use of voiced detectors comes with the added benefit of decreasing the computational time and the complexity of the [AVQA](#) system, at expenses of having to process irrelevant information.

11.3 SUB-TEST D3: TESTING OUT THE PERFORMANCE OF GMM-BASED CLASSIFIERS

The current sub-test explores the usefulness of [GMM](#)-based classifiers in pathological detection tasks. The methodology that is followed is the one introduced in section [10.2.1.3](#).

11.3.1 Trial D3_SusPho

Within this trial three conditions are tested out by varying the type of acoustic material that is employed for training the [UBM](#) and the compensation models in the [IV](#) and [PLDA](#) schemes. Hence, *condition C₁* uses the normophonic registers of the vowel /a/ in the [HUPA](#) database; *condition C₂* uses recordings of normophonic and dysphonic speakers phonating the vowel /a/ in [MEEI](#) (see [10.1.7](#)), [HUPA](#) and [GMar](#); whereas *condition C₃* uses normophonic registers of several partitions including [MEEI](#), [GMar](#), [HUPA](#) and voiced segments of [EUR0M](#) (see [10.1.4](#)). The best results in terms of [AUC](#), [ACC](#) calculated at *Equal Error Rate* ([EER](#)), and related metrics are presented in Table [11.14](#). Likewise, the [DET](#) curve illustrating the best global results are introduced in Figure [11.16](#).

From the obtained outcomes it can be observed that a mere [MAP](#)-adaptation of [UBM](#) models do not provide any further improvement to the performance of the system compared to the baseline [GMM](#), as evidenced by the decreased efficiency of the [GMM-UBM](#) classifier. However, when using more complex systems that stand on the idea of [UBM](#) models some subtle improvements are attained. In particular two considerations can be made. First, the [PLDA](#) provides the best efficiency in terms of [AUC](#). Second, the best results almost always involve the use of normophonic-only registers for training [UBM](#) and compensation models ([MEEI](#), [GMar](#), [HUPA](#) and [EUR0M](#)). Indeed, when the amount of normophonic material is increased there are performance improvements as ascertained by comparing the results of condition *C₁* (which only employs one database of normophonic recordings) to condition *C₃* (which employs 4 databases of normophonic registers).

Table 11.14. *Trial D3_SusPho*: performance of the UBM-based classifiers in the detection of pathologies using the partition of vowel /a/ belonging to the SVD dataset. Results are for the three conditions according to the type of UBM employed. Best outcomes for each type of classifier are in gray, and best absolute result is in red. r_t and r_h stand for the dimensionality of factorial models in IV and PLDA classification respectively.

UBM	Type of classifier	ACC	SP	SE	AUC
-	GMM: MFCC (18)	70.48 ± 2.28	0.67	0.73	0.77
Condition C_1	GMM-UBM: MFCC(16)	69.27 ± 2.37	0.68	0.70	0.76
	IV: MFCC(16) $\rightarrow r_t = 80$	67.15 ± 2.41	0.66	0.68	0.75
	PLDA: MFCC(20) $\rightarrow r_t = 65; r_h = 55$	71.66 ± 2.31	0.70	0.73	0.79
	GMM-SVM: MFCC(16)	71.18 ± 2.32	0.69	0.73	0.77
Condition C_2	GMM-UBM: MFCC(16)	69.20 ± 2.37	0.69	0.69	0.76
	IV: MFCC(16) $\rightarrow r_t = 95$	68.38 ± 2.38	0.66	0.70	0.75
	PLDA: MFCC(16) $\rightarrow r_t = 35; r_h = 40$	72.76 ± 2.28	0.72	0.73	0.79
	GMM-SVM: MFCC(16)	72.01 ± 2.30	0.69	0.74	0.77
Condition C_3	GMM-UBM: MFCC(16)	68.24 ± 2.39	0.67	0.69	0.75
	IV: MFCC(16) $\rightarrow r_t = 35$	70.77 ± 2.33	0.71	0.71	0.78
	PLDA: MFCC(16) $\rightarrow r_t = 50; r_h = 40$	71.32 ± 2.32	0.69	0.73	0.80
	GMM-SVM: MFCC(16)	71.73 ± 2.31	0.70	0.73	0.78

Figure 11.16. *Trial D3_SusPho*: DET curve of the best UBM-based classifiers. The GMM-UBM corresponds to using normophonic registers of HUPA (condition C_1), while the remaining classifiers belong to the test using normophonic recordings of MEEL, GMar, HUPA and EUROM (condition C_3).

11.3.2 Trial D3_RS

The current trial employs the partition based on running speech of the SVD corpus, and different combinations of ancillary databases that define four conditions: condition C_1 , based on normophonic and dysphonic recordings of the sustained phonation of vowels /a/, /i/, /u/ in the HUPA and GMar database next to normophonic voiced segments of the EUROM and Albayzin corpora (see 10.1.5); condition C_2 which considers normophonic-only recordings of the vowels /a/, /i/ and /u/ in the HUPA and GMar database, next

to normophonic voiced segments in sentences of the EUROM and Albayzin databases; *condition* C_3 which characterises the normophonic sentences of MEEI, EUROM and Albayzin; *condition* C_4 which characterises the normophonic sentences of EUROM; and *condition* C_5 which characterises the normophonic sentences of EUROM and Phonedat (see 10.1.6). The best results for each one of the conditions are presented in Table 11.15. Likewise, the DET curve with the best global results are introduced in 11.17.

Table 11.15. *Trial D3_RS*: performance of the UBM-based classifiers in the detection of pathologies using the partition of running speech belonging to the SVD dataset. Results are for the three subtest according to the type of UBM employed. Best outcomes for each type of classifier in gray. r_t and r_h stand for the dimensionality of factorial models in IV and PLDA respectively.

UBM	Type of classifier	ACC	SP	SE	AUC
-	GMM: MFCC (18)	80.32 $\pm$ 2.02	0.74	0.84	0.86
<i>Condition</i> C_1	GMM-UBM: MFCC(16)	76.70 $\pm$ 2.15	0.75	0.78	0.84
	IV: MFCC(14) $\rightarrow r_t = 95$	71.79 $\pm$ 2.29	0.70	0.73	0.79
	PLDA: MFCC(12) $\rightarrow r_t = 65; r_h = 10$	74.01 $\pm$ 2.23	0.74	0.74	0.82
	GMM-SVM: MFCC(10)	77.03 $\pm$ 2.14	0.75	0.78	0.85
<i>Condition</i> C_2	GMM-UBM: MFCC(20)	78.44 $\pm$ 2.09	0.72	0.82	0.86
	IV: MFCC(14) $\rightarrow r_t = 65$	73.00 $\pm$ 2.25	0.71	0.74	0.81
	PLDA: MFCC(10) $\rightarrow r_t = 65; r_h = 25$	75.62 $\pm$ 2.18	0.75	0.76	0.85
	GMM-SVM: MFCC(14)	78.17 $\pm$ 2.10	0.76	0.80	0.85
<i>Condition</i> C_3	GMM-UBM: MFCC(20)	78.44 $\pm$ 2.09	0.72	0.82	0.86
	IV: MFCC(14) $\rightarrow r_t = 65$	73.00 $\pm$ 2.25	0.71	0.74	0.81
	PLDA: MFCC(10) $\rightarrow r_t = 65; r_h = 25$	75.62 $\pm$ 2.18	0.75	0.76	0.85
	GMM-SVM: MFCC(16)	78.71 $\pm$ 2.08	0.76	0.80	0.86
<i>Condition</i> C_4	GMM-UBM: MFCC(20)	76.96 $\pm$ 2.14	0.74	0.79	0.84
	IV: MFCC(12) $\rightarrow r_t = 80$	72.73 $\pm$ 2.26	0.72	0.73	0.81
	PLDA: MFCC(10) $\rightarrow r_t = 65; r_h = 40$	75.62 $\pm$ 2.18	0.74	0.76	0.84
	GMM-SVM: MFCC(16)	77.77 $\pm$ 2.11	0.71	0.81	0.86
<i>Condition</i> C_5	GMM-UBM: MFCC(20)	78.64 $\pm$ 2.08	0.73	0.82	0.85
	IV: MFCC(10) $\rightarrow r_t = 95$	70.65 $\pm$ 2.31	0.69	0.71	0.79
	PLDA: MFCC(12) $\rightarrow r_t = 65; r_h = 2$	76.09 $\pm$ 2.17	0.74	0.77	0.83
	GMM-SVM: MFCC(16)	79.25 $\pm$ 2.06	0.76	0.81	0.86

Despite the best absolute results -as ascertained by the DET curves and the AUC value- are given by the simple GMM classifier, it is worth comparing the influence of the ancillary databases in the performance of UBM-based classifiers. In this respect, *condition* C_3 provides the best performance, indicating that the inclusion of acoustic material based on sustained phonation do not contribute to a performance enhancement (see *condition* C_2). By contrast *condition* C_1 attains, generally, the lowest AUC. This suggests that using normophonic and dysphonic registers for training the UBM and compensation models has not been beneficial for increasing efficiency. *Condition* C_4 is intended to test the behaviour of the system when the auxiliary database

Figure 11.17. *Trial D3_RS*: DET curve of the best UBM-based classifiers. All the classifiers employ normophonic recordings of MEEI, EUROM and Albayzin for training UBM and compensation models.

matches the language of the target partition, while *condition C₅* is similar to *condition C₄* but employing two auxiliary databases. However, no further improvements in efficiency is found in any of these cases.

11.3.3 Discussions on sub-test D₃

During this sub-test several conditions have been considered to explore the performance of systems based on UBM classifiers. In one hand, *trial D3_SusPho* is built with registers relying on sustained phonation for the design of AVQA systems and of UBM and compensation models. The tested conditions allow the examination of interesting scenarios: (i) one where a short amount of normophonic data is employed (*condition C₁*); (ii) other where the amount of normophonic data is increased (*condition C₂*); (iii) and another where normophonic and dysphonic registers are utilised in conjunction (*condition C₃*). Outcomes suggest that models built up using normophonic-only data provides the best results. In addition, and as it might have been expected, increasing the amount of available data enhanced performance. It has also been found that using normophonic and dysphonic registers for training UBM or compensation models do not increase efficiency.

On the other hand, *trial D3_RS* relies on running speech for the design of the AVQA system, whereas the auxiliary databases revolve around the idea of using running speech and/or sustained vowels. As in the previous case, several scenarios are considered: (i) one where normophonic and dysphonic vowels plus voiced segments extracted from speech are employed (*condition C₁*); (ii) other where normophonic-only data of sustained phonation and normophonic voiced segments are used (*condition C₂*); (iii) another using normophonic-only sentences (*condition C₃*); (iv) other using normophonic-only sentences uttered in the same language as SVD (*condition C₄*); (v) and a final which is similar to the latter but increasing the amount of registers used for training (*condition C₅*). Interestingly, the best results are obtained when the simple GMM classifier is employed, indicating that no further improvement

is given by the use of UBM-based classifiers. Not even building models using voiced segments -where supposedly is the most important information for pathology detection- served to improve performance. Two possible explanations for this behaviour include the differences between the acoustic material of the SVD database and that of the auxiliary corpora; or the limited number of available recordings in the ancillary partitions. In this regard, it is worth noting that the SVD dataset is composed of a single sentence that is uttered by German speakers, whereas Albayzin uses diverse sentences produced by Spanish speakers and MEEI contains registers of a text uttered in English. This mismatch might have affected the results as phonetics differs among languages. In an attempt to examine an scenario on which both ancillary and target databases share the same language, *condition C₄* and *condition C₅* have been included. Notwithstanding, results indicate that not even under this setting the performance improved. In this particular case, we might attribute the diminished performance to the mismatch between the acoustic content of SVD, and that of EUROM and/or PhoneDat. A premise in the speaker recognition field is that when the lexical content employed for the ancillary databases matches with the one used for enrolment, a better performance is generally obtained [331]. When there is a mismatch, it is often necessary to include more data to generate more robust models that operate under a text-independent scenario. This is the reason for which *trial D₃_SusPho* reported positive outcomes. During this trial, the ancillary database is composed of sustained vowels, matching with the lexical information of the SVD partition. Moreover, since sustained phonation is relatively unaffected by language this effect is minimised. In any case, results indicate that a topic that is worth to study is in the analysis of other ancillary databases matching the lexical content of the sentences in SVD for training UBM and compensation models.

Finally, it is worth to remark that a common tendency that is observed in both trials is that better results are often obtained when normophonic data is used to construct models. An hypothesis explaining this, might be related to normophonic phonation having a more compact representation in the feature space (as there does not exist variability due to pathologies), which in turn permits a better adjustment of the initialisation models in the UBM systems. It is still necessary, though, to study the influence of dysphonic recordings in the training of compensation models. Some tests -not included in this thesis- were performed under this scenario during *trial D₃_SusPho*, with some negative results. However, it was not possible to perform this type of experiments using running speech due to the lack of available databases, for which this remains an open issue.

11.4 SUB-TEST D4: COMBINATION OF THE BEST SYSTEMS.

The current sub-test is aimed at designing AVQA systems with the knowledge obtained from the previous trials. Methodologically, it follows the stages explained in section 10.2.1.4. The idea is to consider the hierarchical models of *sub-test D₁*, employing the best set of features in *sub-test D₂*

and the UBM-based systems of *sub-test D3*. The acoustic material of different sources is then fused using a logistic regression.

Firstly, a strategy based on feature selection is followed to rank the most consistent and generalist set of features via MIM, mRMR, JMI and a scoring procedure. To this end, the HUPA, GMar and SVD databases are re-parametrised to ensure that each set of features produces the same number of windowed frames. Then, and for each database, the characteristics are concatenated and the three feature selection algorithms are applied. Following a scoring procedure as the one illustrated in Figure 10.14, decisions are generalised.

The top-10 features for each database after having used the aforementioned scoring techniques is presented in Table 11.16. Following the same procedure but for all the databases and all the acoustic material, the resulting global top-10 most generalist and consistent set of characteristics is introduced in Table 11.17.

Table 11.16. *Sub-test D4*: top-ranked features from each database after having applied the scoring procedure of Figure 10.14.

HUPA	SVD	GMar
GNE	CHNR	RALA
PE	GNE	GNE
RALA	RALA	CPPS
CHNR	PE	CHNR
MSP ₂₅	DFA	h^e
LLE	h^e	LLE
CPPS	MSP ₉₅	CIL
MSP ₉₅	LLE	GSampEn
MSH	MSH	DFA
ApEn	GSampEn	PE

Table 11.17. *Sub-test D4*: top-10 most consistent and generalist features according to the feature ranking procedure for all the databases and all the acoustic material. From left to right, the most to the least relevant features.

Rank									
GNE	CHNR	RALA	PE	CPPS	LLE	MSP ₉₅	DFA	h^e	CIL

By comparing results of Tables 11.17 and 11.16 it can be observed that GNE, CHNR and RALA are the most consistent features among databases and sets of characteristics. PE and CPPS are also proficient, but in the case of the first it is just in the 10-th position when considering the GMar database, whereas the latter is not even included in the ranking of the SVD corpus. As a result, the proposed AVQA is to be designed by using GNE, CHNR and RALA for characterisation and GMM classifiers to provide decision scores. In accordance to the results in *sub-test D1*, the proposed systems should consider the sex of the speakers in a hierarchical-like categorisation procedure. Finally, and to include the information of systems based on UBM classifiers, a fusion stage,

at the score level, is to be used to combine the information of the most consistent features and the one provided by the UBM-based classifiers. Regarding the latter, and when using sustained phonation, PLDA systems trained with normophonic data are employed (condition C₃ in trial D₃_SusPho, as in section 11.3.1). For the case of running speech, the GMM algorithm is employed as it has been shown to provide better results than more complex schemes. The whole procedure is summarised graphically in Figure 11.18.

Figure 11.18. Sub-test D₄: resulting methodology after having considered the outcomes of sub-test D₁, sub-test D₂ and sub-test D₃.

11.4.1 Trial D₄_SVD

The results after considering the methodology in Figure 11.18, using the vowels /a/, /i/, /u/, and the sentence in the SVD database are introduced in Table 11.18, whereas the best DET curves are in Figure 11.19. It is worth noting that the results for the vowels are after having fused the information of the system based on GMM and the one based on PLDA.

Table 11.18. Trial D₄_SVD: best results using all the acoustic material in SVD. R.S.: Running speech

	ACC	SP	SE	AUC
Vowel /a/	75.42 ± 2.19	0.74	0.76	0.84
Vowel /i/	72.33 ± 2.27	0.70	0.73	0.81
Vowel /u/	72.20 ± 2.28	0.71	0.73	0.80
R.S.	80.66 ± 2.01	0.70	0.86	0.87
All speech	81.40 ± 1.98	0.71	0.87	0.88

Figure 11.19. Trial D₄_SVD: DET curve for the best results in SVD.

The outcomes indicate that the fusion of different speech material increases performance, achieving in the best case scenario an AUC of 0.88 and a DET curve that is better than the remaining in all operation points.

11.4.2 Trial $D_4_SVD_CrossDatabase$

The current trial is focused on training an [AVQA](#) system with data of the [SVD](#) database and the methodology in Figure [11.18](#), but following a cross-database scenario tested in other four partitions: (i) vowels /a/, /i/ and /u/ of [GMar](#); (ii) the vowel /a/ in [HUPA](#); (iii) registers of the vowel /a/ in [ATIC](#) (see [10.1.9](#)); (iv) and recordings of the vowel /a/ in [DN](#) (see [10.1.8](#)). The best results for all the tested partitions are presented in Table [11.19](#).

Table 11.19. Trial $D_4_SVD_CrossDatabase$: results after using the methodology presented in trial D_4_SVD in a cross-database scenario.

Database	Vowel	ACC	SP	SE	AUC
GMar	/a/	72.00 ± 7.19	0.60	0.83	0.82
	/i/	62.67 ± 7.74	0.47	0.77	0.75
	/u/	72.67 ± 7.13	0.71	0.74	0.79
	All	74.00 ± 7.02	0.64	0.83	0.82
HUPA	/a/	74.46 ± 4.43	0.65	0.85	0.87
ATIC	/a/	78.21 ± 9.16	0.90	0.74	0.93
DN	/a/	82.87 ± 5.49	0.76	0.89	0.94

In general terms the results are acceptable, with [AUC](#) varying between 0.75 to 0.94. When the vowel /a/ is considered [AUC](#) is always superior to 0.82. In particular for the [GMar](#) database, no apparent performance enhancement is obtained when all the vowels are fused.

11.4.3 Discussions on sub-test D_4

The current subtest has been devoted to the design of an [AVQA](#) system with all the knowledge that has been garnered from previous sub-tests. As a result, two sex-independent systems are constructed. The first is based on the most consistent set of features, i. e., [GNE](#), [CHNR](#), [RALA](#) and [GMM](#) classifiers; whereas the second relies on [MFCC](#) and the best [GMM](#)-based classifier in *sub-test* D_3 , i. e., [PLDA](#) for sustained phonation and [GMM](#) for running speech. Results indicate that the best performance is encountered when the scores of the aforementioned systems are fused. In addition, a further improvement is accomplished when systems using vowels and running speech are fused as well (as ascertained in the [SVD](#) database). Even though the improvement is not large in comparison to using running speech alone, it signals a tendency towards the complementarity of the sustained phonation and running speech tasks. When the obtained results are compared to the trials in *sub-test* D_2 some interesting observations can be drawn. First, and in all cases, there is a performance enhancement when the system proposed in this sub-test is employed. The best result in *trial* D_2_SVD-A is of [AUC](#)=0.78, which contrasts with the outcomes in the current sub-test ([AUC](#)=0.84). Similarly for *trial* D_2_SVD-I , the best efficiency is of [AUC](#)=0.74 which is outperformed by the result of *trial* D_4_SVD ([AUC](#)=0.81); whereas for *trial* D_2_SVD-U , the

same tendency is observed (AUC of 0.71 vs. AUC=0.80 in the present sub-test). By contrast, results for running speech are in the same order of magnitude (AUC=0.86 for *trial D2_SVD-RS* and AUC=0.87 in *trial D4_SVD*).

In reference to *trial D4_SVD_CrossDatabase*, some acceptable results are obtained in all cases, having AUC varying from 0.75 to 0.94. It is worth noting that ACC tends to be much lower of what it might have been expected according to the obtained AUC. This might be explained by having used the optimal operation point of the SVD database (for the calculation of ACC) but then testing in a different corpus. AUC is not dependent on any particular operation point and is therefore a better measure of the actual performance of the system. In general terms the results of this cross-database experiment are not that far from the previously reported outcomes, suggesting, that to some extent, the discrepancies between training and testing instances are compensated by the employment of a more robust approach. Indeed, for the GMar database and the vowel /a/, the achieved results in the cross-database scenario (AUC=0.82) are comparable to those of *trial D1_GMar_Sex* (AUC=0.78) and to the best set of features in *trial D2_GMar-A* (AUC=0.84). For GMar and vowel /i/ the best achieved result in *trial D4_SVD_CrossDatabase* is of AUC=0.75, whereas in *trial D2_GMar-I* an AUC of 0.76 has been reported. The same occurs for the vowel /u/, where the best result of the current sub-test is of AUC=0.79 in comparison to an AUC=0.74 achieved in *trial D2_GMar-U*. In reference to the HUPA database, the same tendency is also observed, i. e., AUC for the current trial is of 0.87 whereas in *trial D2_HUPA* it is of 0.85, and in *trial D1_HUPA_sex* it is of 0.81. Finally, the good performance that is observed using both ATIC and DN partitions, confirms the robustness of the system, having achieved AUC of 0.93 and 0.94 respectively.

11.5 GENERAL DISCUSSION

An analysis about factors affecting the design of automatic voice pathology detection systems has been presented in the current section. Extralinguistics such as sex and age have been studied in hierarchical-like schemes, along with different sets of features in a variety of databases containing speech material of diverse type. Moreover, techniques relying on the idea of UBM classifiers have been explored, utilising acoustic material based on sustained phonation and running speech. Results indicate that a segmentation according to the sex of the speaker improves the performance, which in absolute terms varies between 0.3% to 4% depending on the dataset. Results are in line with those found in literature, where the classification accuracy of an automatic detector of pathology is lightly improved by using a manual segmentation of the database according to the sex of the speakers [235]. Respecting the inclusion of the age of the speaker, care should be taken as testing has been performed with a dataset restricted in size, and after having taken strong assumptions regarding the limits that define the frontier between age groups. Regardless of that, results suggest that partitioning the database according to a certain extralinguistic criterion improves performance, indicating the usefulness of hierarchical systems that decompose the

voice pathology detection problem into smaller sub-problems. Indeed, it has been found that the best operation points for female and male systems differ. This is expected as the vocal tract between both sexes differ, which in turn produces subsequent changes in spectral characteristics that consequently impacts the AVQA systems. Accounting for these differences separately simplifies the detection problem and provides performance enhancements.

With respect to the trials involving different sets of features, the outcomes have demonstrated that no single measurement characterises entirely the phenomena related to dysphonia. This suggests that multidimensional approaches should be followed when AVQA systems are designed. Results also indicate that measures based on perturbation and modulation spectra produce positive results in detection of pathologies. Indeed, as it has been ascertained during *sub-test D4*, the most consistent features are GNE, RALA and CHNR (along with CPPS and PE). The feasibility of these features has also been extensively demonstrated by the individual classification they provide and by acceptable correlation observed during experimentation. Interestingly, these features come from different contexts. However, a performance improvement is generally obtained when they work in conjunction, evidencing its complementarity.

In reference to the sub-test involving UBM-based systems, outcomes indicate that this type of classification systems perform well when sustained phonation are used along with ancillary data of normophonic sustained phonation. By contrast, there is not a performance enhancement in any of the tested conditions in the trial involving running speech. Therefore the baseline GMM achieves better outcomes than when using any of the UBM-based systems. This behaviour might be explained by differences among the acoustic material of the databases, and the lack of representative ancillary speech that resembles lexically the content of the sentences in the SVD database. As a conclusion, further testing is needed as the ancillary data is limited, the influence of pathological registers in the training of compensation models for running speech task should also be studied, and phenomena such as the phonetic differences among languages, factors such coarticulation or the lexical divergences between databases might be introducing variability that is hindering system performance. This also highlights the need of having more data available for the study of AVQA systems. Hence, translating methodologies used in speaker recognition (such as those based on UBM classifiers), should start by considering a translation of equivalent databases to the AVQA field. Finally, it is worth noting that extra tests -not presented in this thesis- were performed. These included variations in the manner on how the acoustic material is employed for training the models, e.g., training the UBM with normophonic data and the compensation models with normophonic and dysphonic data, etc. Moreover, other techniques such as *maximum likelihood linear regression*, JFA, ISC, GMM-SVM with *nuisance attribute projections*, z-norms and t-norms were also tested out. However, these results are not included in this thesis as in most of cases the computational burden boosted considerably with no further efficiency enhancement (or even a reduction compared to the simple GMM system).

The best global results for the [SVD](#) corpus are obtained when the information of all the vowels and of the running speech task are fused. In this scenario the performance metrics are as follow: [AUC](#)=0.88 and [ACC](#)=81%. Comparing this results to other works in literature is difficult, as in most cases there exist large differences in the partitions that are used or in the techniques that are employed to assess results. Despite that, there exist two papers that resemble the most the approach presented in this thesis. In the first, [\[332\]](#), [MFCC](#) parameters are used along with perturbation characteristics in a partition of the [SVD](#) dataset containing 1320 pathological and 650 normophonic speakers. Recordings of the vowels /a/, /i/, /u/ produced at normal, high, low and low-high-low pitch are considered for each speaker. When employing single vowels the best [AUC](#) is of 0.74 for vowel /a/ at normal pitch, 0.70 for vowel /i/ at normal pitch, and 0.70 for vowel /u/ at normal pitch. However, when fusing the different intonations of each one of the vowels the efficiency improves to approximately 0.79. Moreover, when the three vowels at the four different intonations are fused, the best result is achieved ([AUC](#)=0.88). However, care must be taken since as stated by the authors, "*these results are optimal in terms of fusion since the fusion parameters have been trained on the test data*". This indicates a certain bias in the logistic regressor procedure which might have affected the performance reported in the paper. In a parallel work described in [\[333\]](#), the same authors compared the performance of the above-mentioned system when using a score-level and a audio-level fusion stage. The best results are obtained in the score-level scenario ([AUC](#)=0.89), with a relative performance improvement of 12% compared to the audio-level case. However, it is important to highlight that when training the algorithms, the whole acoustic material has been employed for training models and then used for evaluation purposes. This implies that results might be biased as well. A fair comparison of the outcomes presented in these two papers and the ones given by our proposed system is not possible except for the case where the vowels /a/, /i/ and /u/ at normal pitch are considered. In this case, the system in [\[332\]](#) achieves [AUC](#)=0.74 for vowel /a/ at normal pitch, 0.70 for vowel /i/ at normal pitch, and 0.70 for vowel /u/ at normal pitch. By contrast the proposed system in this thesis achieves [AUC](#)=0.84 for vowel /a/, 0.81 for vowel /i/ and 0.80 for vowel /u/. Despite the differences in the approaches it can be observed that our system achieves, for each individual vowel, a performance that is better than the ones provided in [\[332\]](#) and a global performance ([AUC](#)=0.88) that is similar to their best system under a more realistic -non biased- scenario.

Despite the differences in the partitions used for experimentation, other related approach is presented in [\[88\]](#), where a subset of the [SVD](#) database composed of 266 normophonic and 263 dysphonic registers (comprising three pathologies: cysts, paralysis and polyps) report an [ACC](#) of 91%. The features are based on correlation and lags of filter-banks calculated after having used a time-frequency decomposition. In a different approach, 101 pathological recordings corresponding to three pathologies and 60 normophonic voices of the [SVD](#) database [\[200\]](#) are used. This paper also employs the [MEEI](#) corpus (113 pathological and 53 normophonic registers) and an Arabian database of

spoken digits (78 pathological and 87 normophonic) with dysphonic recordings belonging to 5 distinct pathologies. By using MFCC features and GMM classifiers, the best ACC for SVD is 80%. In a cross-database experiment the SVD partition is used to build an AVQA system, and then results are assessed in the MEEI and Arabian databases. For the first case ACC reaches up to 81%, whereas for the second a 69% is obtained. Despite the positive outcomes, the authors criticise the ability of conventional characteristics (like MFCC) to accurately detect pathology, but no discussion is given in terms of AUC or in how non-optimal operation points (e. g., EER optimised for the SVD database) might be affecting the reported results. A related work using the same corpora and the same criterion for partitioning the databases is followed in [137] (although the number of recordings differ). By using interlaced derivative patterns of the glottal excitation signal and a SVM classifier, an ACC of 93% is obtained for SVD. In cross-database trials, an ACC of 85% is achieved when testing is performed in MEEI, and of 88% when tested in the Arabic database. Authors in [233] follow the same line of thought (the same database and the same pathologies are considered). However, the MDVP software is used to extract features and a SVM to provide decisions. In this case, the best accuracy for the SVD database is 99% with a subset of features ranked according to the Fisher's discriminant ratio. In a cross-database scenario, results reach up to 70% when tested in the MEEI database, and 71% in the Arabic corpus. Even though it is not possible to compare our results to those of the previous works due to the differences in the employed acoustic material, they serve to illustrate to which extent results are affected when a cross-database approach is followed. In all cases the results reported in literature are in the same line as the ones provided by our system. In particular, we found that in our system AUC varied from 0.75 to 0.94 and that ACC varied between 72% to 84% depending on the tested database. These outcomes evidence the robustness of the approach in providing decision systems about the presence of vocal pathologies. Indeed, the diminished ACC (according to the AUC values that are obtained) might be explained by having trained the threshold used in the likelihood decision function with the SVD database, but then having employed this value for taking decisions in GMar, HUPA, DN and ATIC.

RESULTS ON VOICE PATHOLOGY ASSESSMENT

This chapter is devoted to the experiments investigating the automatic assessment of voice pathologies. As previously introduced during section 10.2.2, four sub-tests are to be performed in this chapter:

- *Sub-test A1*: is designed to test the influence of different sets of characteristics in assessment tasks (see section 10.2.2.1). Since trials are carried out in databases containing registers of the vowel /a/, the same sets of features employed for the characterisation of sustained vowels in detection experiments are considered (see Table 10.1).
- *Sub-test A2*: feature selection methods are used to rank the most relevant features for predicting $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ (see section 10.2.2.2). The idea is to choose a consistent subset that provides the best generalisation results among all the databases. Ordinal classification is employed to predict the traits while accounting for the ordering of the GRBAS evaluations. Moreover, a Gaussian regressor is utilised to deem for the continuous nature of the assessment task. A partition of the SVD dataset, the DN and the ATIC corpora are used for evaluation of results.
- *Subtest A3*: the influence of sex in automatic assessment tasks is studied in this sub-test. To this end, two systems are constructed: one trained with the evaluations of female speakers and the other with the evaluation of males. A description of the methodologies followed is presented in section 10.2.2.3.
- *Sub-test A4*: developed for a clinical validation of the methodologies in *sub-test A2*. To evaluate the results, a speech pathologist has provided an assessment about the reliability of the labels supplied by the system. Likewise, the expert is asked to provide a blind evaluation of vocal quality, but defining the GRBAS scale in a continuous scale. Metrics of error are computed between the assessments provided by the speech pathologist in the evaluations and the labels given by the system. The methodology of current sub-test is presented in section 10.2.2.4.

12.1 SUB-TEST A1: PREDICTION OF $\mathcal{G}$, $\mathcal{R}$ AND $\mathcal{B}$ USING DIFFERENT SETS OF FEATURES.

This sub-test, described in section 10.2.2.1, is composed of three trials that examine the ability that different sets of features provide for the assessment

of vocal pathologies. Each trial is identical except for the dataset that is utilised. Since all databases have been assessed by the same speech therapist, a correlation analysis between pairs of traits is firstly provided in Table 12.1. This serves as an exploratory analysis that might indicate how closely related certain traits are in the different datasets; or to which extent the speech therapist tends to associate a certain trait to another.

Table 12.1. *Sub-test A1*: correlation between pairs of traits for the different databases.

	$\mathcal{G}-\mathcal{B}$		$\mathcal{G}-\mathcal{R}$		$\mathcal{B}-\mathcal{R}$	
	ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$
HUPA	0.74	0.73	0.81	0.79	0.71	0.70
SVD	0.76	0.79	0.80	0.82	0.58	0.60
GMar	0.67	0.67	0.86	0.84	0.53	0.49

As observed, there is a large correlation between $\mathcal{G}-\mathcal{R}$ and $\mathcal{G}-\mathcal{B}$. This is expected as $\mathcal{G}$ is often considered a superclass encompassing $\mathcal{B}$ and $\mathcal{R}$. However, there is a considerable correlation between $\mathcal{B}-\mathcal{R}$ that might be explained, partially, by the presence of pathologies influencing these two traits simultaneously. In any case, this value is surprisingly high, specially for the HUPA database where $\rho_{\mathcal{R}} = 0.70$ and SVD where $\rho_{\mathcal{R}} = 0.60$.

12.1.1 Trial A1_HUPA

The current trial employs registers of the vowel /a/ belonging to the HUPA database. A correlation analysis between features and traits is reported in Table 12.2. Additionally, Table 12.3 reports the results of an ordinal classification that has been carried out with the aim of evaluating the performance of several sets of features in automatic assessment tasks.

Results suggest that the *Pert* set has the highest interrelationship among all features, with characteristics such as *GNE* presenting correlation values superior to 0.38 no matter the trait that is analysed. The *Pert* set also provides one of the best efficiencies when predicting $\mathcal{G}$ ($AMAE=0.37$). Respecting the *MSs* set, it is noteworthy the consistent level of correlation of all features, with values up to 0.37 when using *RALA*. In terms of classification performance, this set provides some of the best outcomes when predicting $\mathcal{G}$ ($AMAE=0.37$) and $\mathcal{R}$ ($AMAE=0.38$). Another well performing set is *SpecCeps*, with features such as *CPPS* having correlations above 0.33 for any trait. Despite that, using *CPPS* in combination with *LHr* decreases results in the ordinal classification procedure. This might be expected due to the low correlation that *LHr* presents and that in some cases does not surpass 0.05. *PLP* provides the best results in classification for $\mathcal{B}$, having $AMAE=0.31$. Finally, and regarding the complexity features, it can be noticed that within the U_e set *LLE* provides the highest level of correlation (above 0.29 for any trait), closely followed by *PE* (above 0.28 for any trait) and *ApEn* (above 0.22 for any trait). By contrast, the highest correlation in the NU_e set is found when employing *RPDE* and *ApEn*. In

Table 12.2. *Trial A1_HUPA*: correlation analysis of the tested features and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits, using registers of the vowel /a/ in the HUPA dataset.

Set	Features	$\mathcal{G}$		$\mathcal{R}$		$\mathcal{B}$	
		ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$
Pert	NNE	-0.01	0.33	-0.01	0.33	-0.03	0.28
	CHNR	-0.52	0.48	-0.51	0.47	-0.42	0.40
	GNE	-0.55	0.54	-0.49	0.48	-0.38	0.40
SpecCeps	PLP	-	0.27	-	0.25	-	0.23
	MFCC	-	0.23	-	0.22	-	0.22
	CPPS	-0.36	0.33	-0.37	0.32	-0.38	0.33
	LHr	-0.03	0.11	-0.07	0.14	-0.05	0.11
MSs	MSH	0.17	0.19	0.17	0.17	0.17	0.18
	CIL	-0.20	0.22	-0.18	0.17	-0.21	0.20
	RALA	-0.36	0.37	-0.32	0.33	-0.32	0.33
	MSP ₂₅	-0.23	0.25	-0.20	0.21	-0.20	0.22
	MSP ₇₅	-0.20	0.21	-0.17	0.18	-0.19	0.20
	MSP ₉₅	-0.21	0.21	-0.18	0.18	-0.20	0.20
U_e	h^e	0.02	0.08	0.03	0.06	-0.08	0.10
	DFA	0.06	0.09	0.08	0.08	-0.07	0.10
	D_2	0.14	0.20	0.11	0.16	0.05	0.15
	LLE	0.32	0.29	0.34	0.32	0.32	0.29
	RPDE	0.13	0.21	0.21	0.25	0.16	0.19
	ApEn	-0.27	0.25	-0.22	0.22	-0.28	0.26
	SampEn	0.11	0.17	0.06	0.14	0.07	0.14
	mSampEn	-0.08	0.13	-0.13	0.14	-0.16	0.16
	GSampEn	-0.03	0.07	0.03	0.08	0.03	0.06
	FuzzyEn	0.08	0.13	0.07	0.13	0.09	0.12
	rHMMEn	0.16	0.19	0.12	0.14	0.06	0.14
	sHMMEn	0.15	0.17	0.12	0.13	0.06	0.13
	PE	0.28	0.29	0.32	0.31	0.30	0.28
NU_e	D_2	0.12	0.12	0.09	0.12	0.09	0.09
	LLE	0.10	0.10	0.11	0.11	0.11	0.10
	RPDE	0.15	0.17	0.20	0.19	0.08	0.11
	ApEn	0.22	0.19	0.19	0.18	0.20	0.17
	SampEn	0.17	0.15	0.14	0.13	0.16	0.13
	mSampEn	-0.04	0.06	-0.06	0.06	-0.03	0.05
	GSampEn	0.09	0.09	0.07	0.09	0.08	0.06
	FuzzyEn	0.14	0.12	0.11	0.12	0.13	0.10
	rHMMEn	0.20	0.19	0.12	0.12	0.13	0.14
	sHMMEn	0.19	0.18	0.12	0.12	0.14	0.14
	PE	0.02	0.09	0.01	0.11	0.00	0.08

general terms, it seems that the uniform technique provides better correlation than its non-uniform counterpart. Notwithstanding, when considering ordinal classification, it seems that neither U_e nor NU_e excel in performance.

12.1.2 *Trial A1_SVD*

The current trial employs recordings of the vowel /a/ belonging to the SVD database. The correlation analysis between the tested features and $\mathcal{G}$, $\mathcal{R}$

Table 12.3. *Trial A1_HUPA*: ordinal classification performance metrics of the tested features and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits, using registers of the vowel /a/ in the HUPA dataset.

Sets	Subset	Classifier	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
			f_β	MAE	AMAE	f_β	MAE	AMAE	f_β	MAE	AMAE
Pert	-	POM	0.63	0.39	0.37	0.66	0.41	0.34	0.57	0.47	0.43
		ELMOP	0.63	0.40	0.37	0.65	0.42	0.35	0.57	0.48	0.43
CPPS+LHr	-	POM	0.51	0.55	0.49	0.65	0.43	0.35	0.51	0.59	0.49
		ELMOP	0.56	0.58	0.44	0.66	0.45	0.34	0.55	0.63	0.45
MFCC	-	POM	0.59	0.47	0.41	0.65	0.42	0.35	0.54	0.55	0.46
		ELMOP	0.59	0.46	0.41	0.68	0.40	0.32	0.56	0.53	0.44
PLP	-	POM	0.61	0.45	0.39	0.67	0.41	0.33	0.53	0.54	0.47
		ELMOP	0.62	0.45	0.38	0.69	0.38	0.31	0.57	0.50	0.43
MSs	-	POM	0.63	0.40	0.37	0.66	0.41	0.34	0.59	0.45	0.41
		ELMOP	0.63	0.41	0.37	0.67	0.40	0.33	0.62	0.45	0.38
U_e	LongRange	POM	0.49	0.67	0.51	0.65	0.49	0.35	0.48	0.69	0.52
		ELMOP	0.49	0.66	0.51	0.63	0.51	0.37	0.45	0.73	0.55
	DynInv	POM	0.57	0.51	0.43	0.67	0.43	0.33	0.55	0.53	0.45
		ELMOP	0.60	0.48	0.40	0.67	0.42	0.33	0.55	0.55	0.45
	Reg	POM	0.52	0.54	0.48	0.62	0.48	0.38	0.49	0.61	0.51
		ELMOP	0.56	0.49	0.44	0.64	0.46	0.36	0.55	0.55	0.45
	EntAtt	POM	0.55	0.51	0.45	0.67	0.41	0.33	0.51	0.56	0.49
		ELMOP	0.60	0.45	0.40	0.64	0.44	0.36	0.59	0.48	0.41
	DynInv	POM	0.54	0.53	0.46	0.66	0.45	0.34	0.50	0.59	0.50
		ELMOP	0.62	0.46	0.38	0.66	0.42	0.34	0.59	0.52	0.41
	Reg	POM	0.52	0.55	0.48	0.65	0.43	0.35	0.51	0.59	0.49
		ELMOP	0.58	0.47	0.42	0.66	0.41	0.34	0.58	0.51	0.42
NU_e	EntAtt	POM	0.52	0.56	0.48	0.65	0.48	0.35	0.50	0.65	0.50
		ELMOP	0.58	0.49	0.42	0.66	0.45	0.34	0.55	0.57	0.45

and $\mathcal{B}$ is presented in Table 12.4. Likewise, results for ordinal classification are shown in Table 12.5.

Similarly to *Trial A1_HUPA*, the best results are achieved when using the *Pert* set (with correlation values above 0.53 using *CHNR* and above 0.44 using *GNE*) and classification results that are consistently higher than in the remaining sets (*AMAE* for $\mathcal{G}$ =0.60, $\mathcal{B}$ =0.74 and $\mathcal{R}$ =0.78). Considering the *MSs* set, *RALA* is once again the best correlated characteristic, although, being closely followed by the remaining (in all cases always above 0.4) but for *MSH* (above 0.18). The performance of the *MSs* set is one of the highest among the trial (*AMAE* ranges from 0.69 in $\mathcal{G}$ to 0.84 in $\mathcal{R}$). With regards to *SpecCeps*, *CPPS* and *PLP* present the highest correlation values within the set (always superior to 0.31 for all traits), but its performance in ordinal classification is moderate (specially for *CPPS+LHr* where *AMAE* ranges from 0.92 to 1.13). Respecting the U_e set, it can be noticed that *PE* and *LLE* achieve levels of correlation above 0.41 for all traits. By contrast, the regularity features (except

Table 12.4. *Trial A1_SVD*: correlation analysis of the tested features and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits, using registers of the vowel /a/ in the *SVD* dataset.

Set	Features	$\mathcal{G}$		$\mathcal{R}$		$\mathcal{B}$	
		ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$
Pert	NNE	0.02	0.45	0.04	0.43	0.02	0.43
	CHNR	-0.62	0.58	-0.55	0.53	-0.58	0.55
	GNE	-0.57	0.54	-0.46	0.44	-0.59	0.57
SpecCeps	PLP	-	0.31	-	0.31	-	0.30
	MFCC	-	0.28	-	0.20	-	0.25
	CPPS	-0.34	0.32	-0.32	0.30	-0.34	0.32
	LHr	-0.16	0.23	-0.20	0.27	-0.20	0.26
MSs	MSH	-0.24	0.23	-0.19	0.18	-0.23	0.23
	CIL	0.49	0.45	0.47	0.44	0.48	0.44
	RALA	0.55	0.51	0.51	0.48	0.54	0.51
	MSP ₂₅	0.45	0.42	0.42	0.40	0.47	0.43
	MSP ₇₅	0.48	0.44	0.45	0.43	0.48	0.44
	MSP ₉₅	0.52	0.48	0.50	0.48	0.55	0.50
U_e	h^e	0.29	0.31	0.21	0.22	0.24	0.28
	DFA	0.32	0.33	0.24	0.24	0.26	0.28
	D_2	0.01	0.30	0.00	0.22	0.01	0.29
	LLE	0.44	0.41	0.42	0.41	0.47	0.42
	RPDE	0.30	0.33	0.39	0.40	0.27	0.30
	ApEn	-0.18	0.17	-0.12	0.13	-0.22	0.20
	SampEn	0.06	0.14	0.05	0.09	0.09	0.14
	mSampEn	-0.11	0.12	-0.16	0.15	-0.07	0.09
	GSampEn	-0.02	0.13	0.05	0.12	-0.02	0.13
	FuzzyEn	-0.02	0.16	0.05	0.09	0.03	0.14
	rHMMEEn	0.12	0.17	0.03	0.11	0.13	0.16
	sHMMEEn	0.14	0.17	0.08	0.13	0.15	0.17
	PE	0.49	0.43	0.53	0.48	0.48	0.42
NU_e	D_2	0.16	0.13	0.18	0.13	0.22	0.16
	LLE	0.00	0.08	-0.04	0.10	-0.02	0.09
	RPDE	0.27	0.26	0.26	0.24	0.21	0.21
	ApEn	0.14	0.18	0.16	0.15	0.22	0.21
	SampEn	0.12	0.13	0.11	0.11	0.19	0.16
	mSampEn	-0.03	0.12	0.00	0.08	0.04	0.09
	GSampEn	0.12	0.13	0.14	0.12	0.18	0.16
	FuzzyEn	0.13	0.16	0.16	0.14	0.20	0.19
	rHMMEEn	0.01	0.04	-0.07	0.06	0.02	0.04
	sHMMEEn	0.00	0.03	-0.06	0.05	0.02	0.03
	PE	0.14	0.13	0.17	0.15	0.17	0.15

ApEn where) seems to be more modestly correlated. In terms of ordinal classification performance, EntAtt (where PE is contained) presents some of the most competitive outcomes next to DynInv (AMAE ranging from 0.72 to 0.97 for these two sets). Respecting NU_e , results are generally worse than those obtained with the U_e set, in terms of AMAE.

12.1.3 Trial A1_GMar

The current trial employs recordings of the vowel /a/ belonging to the GMar database. The correlation analysis between the tested features and the

Table 12.5. *Trial A1_SVD A1*: ordinal classification performance metrics of the tested features and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits, using registers of the vowel /a/ in the *SVD* dataset.

Sets	Subset	Classifier	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
			f_β	MAE	AMAE	f_β	MAE	AMAE	f_β	MAE	AMAE
Pert	-	POM	0.54	0.47	0.60	0.42	0.35	0.74	0.37	0.43	0.78
		ELMOP	0.56	0.47	0.59	0.47	0.36	0.74	0.36	0.44	0.82
CPPS+LHr	-	POM	0.26	0.73	0.92	0.25	0.49	1.18	0.26	0.54	1.13
		ELMOP	0.30	0.79	1.03	0.34	0.51	1.19	0.26	0.57	1.27
MFCC	-	POM	0.37	0.64	0.80	0.35	0.45	0.95	0.34	0.51	0.95
		ELMOP	0.44	0.58	0.73	0.42	0.42	0.89	0.34	0.51	0.96
PLP	-	POM	0.44	0.57	0.70	0.45	0.39	0.78	0.38	0.46	0.83
		ELMOP	0.48	0.54	0.68	0.43	0.40	0.82	0.36	0.46	0.85
MSs	-	POM	0.47	0.57	0.69	0.47	0.38	0.78	0.38	0.46	0.84
		ELMOP	0.47	0.58	0.74	0.44	0.41	0.91	0.34	0.47	0.94
U_e	LongRange	POM	0.30	0.75	1.02	0.24	0.57	1.38	0.24	0.63	1.33
		ELMOP	0.33	0.74	1.04	0.28	0.55	1.31	0.25	0.61	1.29
	DynInv	POM	0.44	0.61	0.75	0.41	0.44	0.97	0.35	0.48	0.91
		ELMOP	0.46	0.57	0.71	0.44	0.42	0.92	0.36	0.46	0.87
	Reg	POM	0.19	0.87	1.06	0.22	0.55	1.35	0.22	0.61	1.32
		ELMOP	0.41	0.64	0.84	0.33	0.51	1.19	0.23	0.60	1.31
	EntAtt	POM	0.43	0.61	0.72	0.34	0.43	0.92	0.37	0.46	0.82
		ELMOP	0.49	0.55	0.69	0.44	0.42	0.89	0.35	0.46	0.87
	DynInv	POM	0.38	0.67	0.84	0.34	0.50	1.15	0.31	0.55	1.15
		ELMOP	0.44	0.61	0.74	0.41	0.44	0.97	0.33	0.49	0.97
	Reg	POM	0.28	0.80	0.97	0.28	0.52	1.23	0.26	0.58	1.19
		ELMOP	0.45	0.63	0.80	0.39	0.47	1.05	0.29	0.54	1.11
NU_e	EntAtt	POM	0.18	0.93	1.19	0.20	0.57	1.43	0.22	0.61	1.34
		ELMOP	0.35	0.72	1.02	0.36	0.50	1.19	0.27	0.56	1.22

$\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits is presented in Table 12.6, whereas the results for the ordinal classification are introduced in Table 12.7.

As in the previous two trials, the largest correlation values are generally obtained with the *Pert* set, having $\rho_{\mathcal{R}} = 0.39$ when using *GNE* and *CHNR*. This set also achieves good performance in ordinal classification tasks, specially for $\mathcal{G}$ (AMAE=0.50) and $\mathcal{R}$ (AMAE=0.51). Relating to the *MSs* set, correlation values are amongst the largest within the trial, ranging from 0.44 with *RALA* to 0.11 with *MSH*. In relation to the *SpecCeps* set, correlations are above 0.23 in all cases, but with *CPPS* and *LHr* providing the highest values. Notwithstanding, the best classification performance is obtained with the *PLP* features, specially for predicting $\mathcal{B}$ (AMAE=0.32). With reference to the U_e set there is a modest ordinal classification performance when examining each one of the subsets. The results of the NU_e set vary a lot. Indeed, no single feature correlated above 0.24. Besides that, the ordinal classification performance of all of its subsets is moderate in comparison to the outcomes of other sets.

Table 12.6. *Trial A1_GMar*: correlation analysis of the tested features and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits, using registers of the vowel /a/ in the *GMar* dataset.

Set	Features	$\mathcal{G}$		$\mathcal{R}$		$\mathcal{B}$	
		ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$
Pert	NNE	-0.02	0.23	0.02	0.21	-0.04	0.20
	CHNR	-0.47	0.45	-0.42	0.40	-0.39	0.39
	GNE	-0.46	0.46	-0.43	0.43	-0.40	0.41
SpecCeps	PLP	-	0.32	-	0.29	-	0.29
	MFCC	-	0.28	-	0.25	-	0.29
	CPPS	-0.39	0.36	-0.25	0.23	-0.35	0.32
	LHr	-0.29	0.35	-0.24	0.30	-0.26	0.32
MSs	MSH	-0.13	0.13	-0.12	0.12	-0.11	0.12
	CIL	0.34	0.31	0.27	0.24	0.32	0.30
	RALA	0.44	0.44	0.41	0.42	0.38	0.39
	MSP ₂₅	0.36	0.33	0.25	0.24	0.35	0.33
	MSP ₇₅	0.34	0.31	0.24	0.22	0.33	0.30
	MSP ₉₅	0.38	0.34	0.28	0.25	0.39	0.36
U_e	h^e	0.06	0.17	0.16	0.21	0.02	0.14
	DFA	0.03	0.12	0.09	0.13	0.00	0.10
	D_2	0.06	0.20	0.07	0.20	0.05	0.18
	LLE	0.31	0.27	0.28	0.25	0.36	0.31
	RPDE	0.09	0.18	0.10	0.18	0.10	0.15
	ApEn	-0.24	0.21	-0.18	0.16	-0.25	0.22
	SampEn	0.07	0.13	0.04	0.10	0.09	0.16
	mSampEn	-0.14	0.15	-0.09	0.10	-0.14	0.13
	GSampEn	-0.05	0.10	-0.03	0.08	-0.05	0.09
	FuzzyEn	0.06	0.10	0.02	0.09	0.15	0.13
	rHMMEn	0.08	0.18	0.07	0.17	0.08	0.16
	sHMMEn	0.11	0.17	0.08	0.16	0.11	0.16
	PE	0.29	0.30	0.23	0.24	0.30	0.28
NU_e	D_2	0.15	0.12	0.08	0.08	0.20	0.16
	LLE	0.00	0.11	0.02	0.11	-0.04	0.11
	RPDE	-0.02	0.05	0.08	0.10	-0.05	0.06
	ApEn	0.16	0.17	0.06	0.10	0.24	0.19
	SampEn	0.14	0.14	0.04	0.08	0.20	0.16
	mSampEn	0.00	0.07	-0.05	0.10	0.04	0.05
	GSampEn	0.15	0.14	0.06	0.07	0.19	0.17
	FuzzyEn	0.19	0.18	0.09	0.10	0.23	0.21
	rHMMEn	0.06	0.06	0.01	0.02	0.05	0.06
	sHMMEn	0.05	0.05	0.01	0.02	0.05	0.05
	PE	0.12	0.12	0.09	0.10	0.17	0.15

12.1.4 Discussions on sub-test A1

The results are fairly consistent in the three databases. From the presented outcomes it can be inferred that the *Pert* set provides the best correlation among sets, as ascertained by the results of the three trials. Likewise, both *MSs* and *SpecCeps* also present some of the largest correlation values and a homogeneous behaviour among trials. By contrast, the complexity sets (U_e and NU_e) provide greatly fluctuating outcomes. In terms of individual characteristics, a common pattern of high correlation and good performance is

Table 12.7. *Trial A1_GMar*: ordinal classification performance metrics of the tested features and the $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$ traits, using registers of the vowel /a/ in the *GMar* dataset.

Sets	Subset	Classifier	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
			f_β	MAE	AMAE	f_β	MAE	AMAE	f_β	MAE	AMAE
Pert	-	POM	0.48	0.66	0.52	0.65	0.49	0.36	0.46	0.65	0.55
		ELMOP	0.51	0.64	0.50	0.66	0.54	0.34	0.50	0.62	0.51
CPPS+LHr	-	POM	0.35	0.76	0.65	0.65	0.50	0.36	0.33	0.76	0.67
		ELMOP	0.17	0.97	0.83	0.64	0.58	0.37	0.31	0.76	0.69
MFCC	-	POM	0.39	0.79	0.61	0.65	0.50	0.35	0.41	0.73	0.59
		ELMOP	0.51	0.68	0.50	0.65	0.47	0.35	0.39	0.79	0.62
PLP	-	POM	0.42	0.72	0.59	0.66	0.46	0.34	0.41	0.73	0.60
		ELMOP	0.49	0.70	0.51	0.68	0.45	0.32	0.45	0.73	0.55
MSs	-	POM	0.45	0.71	0.55	0.67	0.50	0.34	0.40	0.70	0.61
		ELMOP	0.35	0.79	0.66	0.66	0.53	0.34	0.35	0.71	0.66
U_e	LongRange	POM	0.19	0.96	0.82	0.65	0.58	0.35	0.36	0.74	0.64
		ELMOP	0.29	0.93	0.71	0.64	0.59	0.36	0.35	0.77	0.66
	DynInv	POM	0.34	0.82	0.66	0.67	0.51	0.33	0.41	0.70	0.60
		ELMOP	0.37	0.82	0.63	0.67	0.50	0.33	0.42	0.72	0.58
	Reg	POM	0.33	0.83	0.67	0.66	0.53	0.34	0.37	0.72	0.64
		ELMOP	0.38	0.75	0.63	0.67	0.51	0.34	0.39	0.70	0.62
	EntAtt	POM	0.27	0.86	0.73	0.66	0.53	0.34	0.34	0.76	0.67
		ELMOP	0.46	0.66	0.55	0.67	0.49	0.34	0.45	0.65	0.55
	DynInv	POM	0.22	0.93	0.79	0.66	0.56	0.35	0.26	0.83	0.75
		ELMOP	0.35	0.81	0.65	0.64	0.54	0.36	0.38	0.74	0.62
	Reg	POM	0.30	0.86	0.71	0.65	0.54	0.36	0.36	0.74	0.64
		ELMOP	0.43	0.75	0.57	0.64	0.54	0.37	0.45	0.66	0.55
NU_e	EntAtt	POM	0.20	0.91	0.81	0.65	0.56	0.36	0.27	0.79	0.73
		ELMOP	0.23	0.89	0.77	0.65	0.56	0.36	0.33	0.79	0.68

generally observed with *RALA*, *CPPS*, *GNE* and *CHNR*. Nonetheless, other well-performing/well-correlated characteristics include *LLE*, *ApEn* and *PE* (in the U_e set). An interesting observation from the sub-test is that features tend to perform equally well no matter the trait that is analysed. This might contradict the reasoning of some features correlating better with some traits as they are expected to measure a certain phenomenon of pathological voice production. For instance, we would expect characteristics measuring modulation noise (such as *RPDE*) to have a higher correspondence to the $\mathcal{R}$ trait, whereas those indicating the presence of additive noise (such as perturbation measures) to be more interrelated to $\mathcal{B}$. This is because additive noise tends to be associated to the perception of breathiness due to turbulence, whereas modulation noise tends to be interrelated to the roughness perception. However, this is not the case, and in contrast a subset of characteristics dominate over $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$. It is important to remark that features have been evaluated individually, or grouped into subsets according to the domain they have

been computed. Thus, an analysis that assesses the complementarity of features -even if calculated in different domains- in the search of a combination that most reliably characterises vocal quality, should be performed.

To provide further evidences about the performance of each tested sets of characteristics, three filter feature selection algorithms are employed ([MIM](#), [mRMR](#) and [JMI](#)). The aim is to rank the features within each set from the most to the least relevant (according to the filter algorithms). The results of the ranking procedure for the [HUPA](#), [SVD](#) and [GMar](#) databases are displayed in the Appendix, in Tables [B.1](#), [B.2](#) and [B.3](#) respectively. To generalise the ranking results disregarding the effect of the database and the filter algorithm, a scoring procedure, similar to the one followed in Figure [10.14](#) is performed. In this manner the best ranked features receive a minimum score, while the worst performing attain the maximum value. Results are then summed up for the three feature selection techniques and for the three databases. At the end, the features with the lowest scores are acknowledged as the most consistent and best performing within the set. The outcomes of the scoring procedure applied to each one of the feature sets are reported in Table [12.8](#).

Table 12.8. *Subtest A1*: best characteristics within each feature set and among all the databases, according to the filter feature selection procedures.

Set	Trait	Features					
Pert	$\mathcal{G}$	CHNR	GNE	NNE			
	$\mathcal{B}$	CHNR	GNE	NNE			
	$\mathcal{R}$	GNE	CHNR	NNE			
SpecCeps	$\mathcal{G}$	CPPS	MFCC	LHr			
	$\mathcal{B}$	CPPS	MFCC	LHr			
	$\mathcal{R}$	LHr	CPPS	MFCC			
MSs	$\mathcal{G}$	RALA	CIL	MSP ₉₅			
	$\mathcal{B}$	RALA	MSP ₉₅	CIL			
	$\mathcal{R}$	RALA	MSP ₉₅	CIL			
U_e	$\mathcal{G}$	PE	LLE	h^e	DFA	RPDE	
	$\mathcal{B}$	PE	LLE	h^e	DFA	ApEn	
	$\mathcal{R}$	PE	LLE	DFA	D_2	h^e	
NU_e	$\mathcal{G}$	LLE	ApEn	h^e	DFA	PE	
	$\mathcal{B}$	LLE	h^e	PE	ApEn	DFA	
	$\mathcal{R}$	LLE	DFA	h^e	PE	RPDE	

Table [12.8](#) confirms the good performance given by [RALA](#), [CPPS](#), [GNE](#) and [PE](#). It also permits a direct examination of the U_e and NU_e sets, where features such as [LLE](#), [PE](#) and [RPDE](#) are well ranked in both cases (after disregarding long-range estimators such as [DFA](#) and h^e as they are independent of the reconstruction). Since both techniques are measuring the same nonlinear properties, it is of interest to drop one of the sets in favour of the other not to introduce redundancy in further experiments while alleviating some of the computational load. With this in mind, the correlation results (in terms of $\rho_{\mathcal{R}}$) for the best correlated features which are common for both U_e and NU_e ([LLE](#), [PE](#) and [RPDE](#)) are introduced in Table [12.9](#). The outcomes reveal that the U_e characteristics are always better correlated than the features on the NU_e

set. Due to this reason and due to the lack of consistency of NU_e during the three trials of the current sub-set, this set has been disregarded from further testing.

Table 12.9. *Sub-test A1*: $\rho_{\mathcal{R}}$ for the complexity sets U_e and NU_e for the three tested databases and the best performing features according to the feature selection procedure.

Trial	Set	HUPA		SVD		GMar	
		U_e	NU_e	U_e	NU_e	U_e	NU_e
$\mathcal{G}$	LLE	0.29	0.10	0.41	0.08	0.27	0.11
	PE	0.29	0.09	0.43	0.13	0.30	0.11
	RPDE	0.21	0.17	0.33	0.26	0.18	0.11
$\mathcal{R}$	LLE	0.32	0.11	0.41	0.10	0.25	0.12
	PE	0.31	0.11	0.48	0.15	0.24	0.10
	RPDE	0.25	0.19	0.40	0.24	0.18	0.15
$\mathcal{B}$	LLE	0.29	0.10	0.42	0.09	0.31	0.05
	PE	0.28	0.08	0.42	0.15	0.28	0.10
	RPDE	0.19	0.11	0.30	0.21	0.15	0.06

Finally, it is worth noting the good behaviour that the descriptor $\rho_{\mathcal{R}}$ presents, specially with some features that otherwise would be considered irrelevant according to the classical correlation measure ρ . One example is in NNE , which in several cases produced a ρ below 0.1, but much more superior when using $\rho_{\mathcal{R}}$. At the end, the relevance of NNE as assessed by the feature ranking and the ordinal classification, agrees with the values given by $\rho_{\mathcal{R}}$, and thus acknowledges its consistency. An additional advantage of such descriptor, consists on the possibility of measuring the correlation of multidimensional characteristics such as PLP and $MFCC$, giving a general idea about their performance as a whole.

12.2 SUB-TEST A2: GENERALISATION OF RESULTS THROUGH FILTER SELECTION METHODS.

Sub-test A2 examines the best performing and most consistent characteristics for assessment tasks. The methodological stages that are followed have been introduced in section 10.2.2.2. Contrary to the previous experiment on which filter feature selection algorithms are applied to differentiate the most relevant features within a set, the current trials consider all characteristics altogether. Thus, for each database, a *supermatrix* is created by concatenating all sets of features (disregarding NU_e in favour of U_e as discussed previously). Then, the same filter feature selection algorithms and scoring procedures described in *sub-test A1* are applied. The outcomes after following such process are referred in Table 12.10.

It can be noticed that the most informative features (for all traits) are GNE , PE , $RALA$ and $CHNR$, with other features such as some PLP NNE or MSP_{95} performing well. For the sake of interpretability only 3 are utilised, as this reduced set provides interesting information that might not be achieved with

Table 12.10. *Sub-test A2*: top-10 ranked features, after considering the three feature selection techniques and the scoring procedure. Results are for the $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$ traits.

$\mathcal{G}$	$\mathcal{B}$	$\mathcal{R}$
CHNR	CHNR	PE
PE	GNE	GNE
GNE	PE	RALA
RALA	PLP(2)	CHNR
PLP(3)	RALA	PLP(3)
PLP(2)	MSP ₉₅	NNE
MSP ₇₅	PLP(3)	MSP ₉₅
NNE	NNE	CPPS
CPPS	CPPS	LLE
CIL	LLE	MFCC(16)

a larger feature space. For instance, limiting the analysis to just three characteristics permits the construction of 3-dimensional projections that expound the separability offered by the features.

Following this idea, and to enlighten the observation of results in the 3-dimensional space, a per-file statistic is computed as the average value of all the frames belonging to a recording. This concentrates all the information of the register into a single point, thus avoiding dispersion in the feature space that otherwise would be encountered with a per-frame approach. Additionally, and as proven later after extensive experimentation, this averaging process has been useful for improving results in automatic assessment tasks. As a matter of example and to illustrate the dispersion phenomenon, Figure 12.1 introduces the CPPS contour formed after plotting all the frames of two voice recordings, one labelled $\mathcal{G}=0$ (blue) and the other $\mathcal{G}=2$ (red). As observed, the register with label 0 presents a very consistent behaviour, with a high CPPS value which is typical of voices not affected by the presence of turbulence noise. By contrast, the register labelled 2 contains a widespread distribution of points with continuous drops and rises. Indeed, in some utterances the values are practically similar to those attained by the register labelled 0. This indicates that even in pathological conditions, there exist segments resembling normophonic behaviour.

Since the reduced supermatrix formed by three characteristics contains data -in a per-file fashion- of the three databases, a further processing is performed to remove those values that might be considered non-representative of the analysed label and trait. This resembles an outlier-removal procedure, but accounting for the multidimensional nature of the supermatrix. To this end, and in view of the compact clusters of the group of features (see for instance the projections in Figure 12.2), each label (1, 2 or 3) within each trait ($\mathcal{G}$, $\mathcal{B}$ or $\mathcal{R}$) is modelled using a GMM trained with 1 Gaussian. Then, the log-likelihood between each data point and the model to which it belongs is computed. The resulting score measures how likely is the data-

Figure 12.1. *Sub-test A2*: contour formed after plotting the CPPS of two voices of the HUPA dataset. The graphic in red is for a voice with $\mathcal{G}=2$, whereas the one in blue is for one having $\mathcal{G}=0$.

point to be generated by the model. Boxplots of the scores are then drawn, as they provide information about the distribution of the points. Only utterances whose scores are above a certain threshold are kept for further testing, hence providing a more compact representation of the label under study. After some testing using different threshold values, the median or 50-th percentile is used due to the good results obtained. A graphic summarising the whole process is introduced in Figure 12.2. In the same manner, Figure 12.3 presents 3-dimensional plots for each one of the traits, using the per-file scheme and the outlier-removal-like procedure. The graphic has been colour-coded to represent the labels. The three depicted features, in all cases, correspond to RALA, GNE and CHNR. This is due to some intensive testing that demonstrated good performance using this three features exclusively in contraposition to other combinations.

Figure 12.2. *Sub-test A2*: procedure for finding a compact representation of the feature space.

From the projections it can be inferred, that for all traits, the selected features offer a consistent separability between the most extreme cases, as determined by the two very differentiated clusters formed by labels 0 and 3 (specially for $\mathcal{G}$ and $\mathcal{B}$). However, for labels 1 and 2, certain interweaving is observed. Certainly, distinguishing between a grave disorder (3) and normo-phonía (0) is easier than distinguishing between consecutive labels like light (1) or moderate (2) disorder.

Figure 12.3. *Sub-test A2*: representation in a per-file basis of the most relevant features according to the feature ranking procedure: $\mathcal{GNE}$, $\mathcal{CHNR}$ and $\mathcal{RALA}$. Projection for $\mathcal{G}$ is in Figure 12.3a, $\mathcal{B}$ in 12.3b, $\mathcal{R}$ in 12.3c.

Since the projections only provide visual information regarding the separability offered by the features, the next two trials are designed for assessing their validity in automatically predicting voice quality. In order to evaluate the performance of the system in the following trials, the confusion matrices resulting from the ordinal regression procedures are to be depicted. Moreover, and for comparing these results to a certain baseline, an artificial confusion matrix is constructed by assuming the existence of 4 $\mathbf{pmf}$ for each label. Each $\mathbf{pmf}$ is designed -optimistically- to ensure a large probability of occurrence for the label itself and its neighbour labels (0.9), and low elsewhere (the remaining 0.10). A graphic depicting the $\mathbf{pmf}$ for each one of the labels is introduced in Figure 12.4. As a matter of example, the $\mathbf{pmf}$ for label 0 assumes that a sample extracted randomly will produce with a probability of 0.7 a value labelled as 0, with a probability of 0.2 a value labelled 1 and with a probability of 0.05 a value that would be either 2 or 3. Values of the confusion matrices are randomly sampled from these $\mathbf{pmf}$ by means of an *inverse transform sampling* procedure [334]. The number of samples is the same as the number of instances of the target label.

12.2.1 Trial A2_SVD

For assessment purposes a partition of 121 registers of the $\mathcal{SVD}$ dataset is employed as described in section 10.2.2.2. $\mathcal{POM}$ and $\mathcal{ELMOP}$ ordinal classifiers are trained for testing out the ability of the selected features in predicting $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$. The outcomes of the ordinal classification procedure are in-

Figure 12.4. Sub-test A2: pmf for the labels 0, 1, 2 or 3. These are used in the design of the random samples generator.

roduced in Table 12.11. Similarly, the confusion matrices of the technique rendering the best results are presented in Figure 12.5.

Table 12.11. Trial A2_SVD: best results obtained using the ordinal classification techniques and the SVD testing partition.

	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
	f_{β}	MAE	AMAE	f_{β}	MAE	AMAE	f_{β}	MAE	AMAE
POM	0.57	0.48	0.50	0.50	0.56	0.54	0.46	0.64	0.63
ELMOP	0.53	0.59	0.58	0.47	0.60	0.58	0.45	0.67	0.68

As observed from Table 12.11, the POM algorithm achieves the best classification results in terms of AMAE for all traits. In spite of that, the outcomes provided by both classifiers are close to each other suggesting consistency of the selected features and of the techniques. From the confusion matrix it can be inferred that, in general, large errors are encountered between the label 1 and its neighbour label 0. In terms of SE, label 1 just ranges from 19% to 36%. Similarly, the label 2 presents -for all traits- large errors, specially due to confusions with the neighbour 3. By contrast, 0 and 3 are very well represented. In particular, label 0 has SE values ranging from 77.8% to 90.4%, whereas 3 ranges from 53.3% to 71.7%. Compared to the artificial matrix, the results of the system are very much alike for $\mathcal{G}$, while for the remaining traits the proposed system is outperformed in some cases whereas in others the contrary occurs. One interesting observation is that most of the errors are distributed around the diagonal, i. e., errors arise in the neighbourhood of the target label. To provide more hindsight about this particular, a weighted version of AMAE is used to measure the influence of allowing this type of errors in the calculation of performance metrics. The measure is a modification of the statistics extracted from the weighted confusion matrix and discussed in [335], but adapted to the context of ordinal classification. In this respect, the classical AMAE assumes that the weight that is usually associated to errors around diagonal is equal to of 1. The *weighted-AMAE* is a generalisation that allows this value to be varied in the range 0 -where an error around the diagonal is regarded as accurate- to 1 -the usual case-. The graphic that arises from plotting the weighted-AMAE for a range of weights, and for the proposed and the artificial system, is presented in Figure 12.6.

(a) Confusion matrix for the $\mathcal{G}$ trait using POM. (b) Confusion matrix for the $\mathcal{B}$ trait using POM. (c) Confusion matrix for the $\mathcal{R}$ trait using POM.

Figure 12.5. *Trial A2_SVD*: confusion matrices for the ordinal classification (right) and their artificial counterparts (left) using the *SVD* partition. Results for $\mathcal{G}$ are in Figure 12.5a, for $\mathcal{B}$ in 12.5b, and for $\mathcal{R}$ in 12.5c. The values next to the matrix indicate the *SP* and *SE* of the system. Each cell contains the number of incidences on top, and the percentage of occurrence at the bottom.

As observed, by setting the weights to 0 and thus regarding the errors around the diagonal as positive outcomes, *AMAE* is decreased drastically to around 0.15 for $\mathcal{G}$ and $\mathcal{B}$, and to approximately 0.20 for $\mathcal{R}$. This supports the previous observations made in the confusion matrices, where most of the errors seemed to have occurred in the neighbourhood of the diagonal.

Figure 12.6. *Trial A2_SVD*: weighted-AMAE for weights varied between 0 and 1. Results for the *SVD* database.

Now, and since the assessment of voice quality should consider the whole spectrum of possible evaluations ranging from normophonia to severe dysphonia, the use of a discrete scale might not be optimal. Hence, and to acknowledge for the continuous nature of the assessment task, a Gaussian regressor is tested out. The decisions given by the regressor are then compared to the original $\mathcal{G}$, $\mathcal{R}$, $\mathcal{B}$ labels to compute error metrics. Following that approach, Table 12.12 presents the RMSE and MAE values when using the *SVD* partition. In addition, raster plots and the *Probability Distribution Function* (pdf) representing the deviation between predicted and target labels are introduced in Figure 12.7. In particular, the raster plot depicts the occurrences of the predicted label. It is expected that the closer the predictions to the targets, the more compact the distribution of points around it. Likewise, the pdf models the raster plot deviations. In systems closely following the target label the distribution is expected to be unimodal with large peaks centred around it but smooth elsewhere.

Table 12.12. *Trial A2_SVD*: error measures for the Gaussian regressor. Figure (a) presents the global values of RMSE and MAE, whereas (b) presents the RMSE per label within each trait.

	RMSE	MAE		0	1	2	3
$\mathcal{G}$	0.69	0.51	$\mathcal{G}$	0.96	0.61	0.57	0.61
$\mathcal{B}$	0.72	0.57	$\mathcal{B}$	0.63	0.56	0.67	0.96
$\mathcal{R}$	0.76	0.58	$\mathcal{R}$	0.95	0.62	0.58	0.87
(a) RMSE & MAE per class.			(b) RMSE calculated for each trait and label.				

Analysing the results obtained for the $\mathcal{G}$ trait, it can be inferred that the label 3 is well represented as given by the prominent peak in the pdf of the figure, and the raster plot centred around the value 3. Even though there are certain points that deviate, the performance in general terms is good (RMSE=0.69). The behaviour of label 2 is acceptable, having some errors in-between labels 0 and 1, but specially 3. By contrast, there is a larger overlap between the distributions of labels 0 and 1 as observed from both raster and pdf plots. When analysing $\mathcal{B}$, a large overlap between 2 and 3 is noticed,

(a) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{G}$ trait.

(b) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{B}$ trait.

(c) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{R}$ trait.

Figure 12.7. *Trial A2_SVD*: raster plots and pdf modelling the differences between the target label and the label predicted by the system. The deviations are normalised around the target label. Results for $\mathcal{G}$ are in Figure 12.7a, for $\mathcal{B}$ in 12.7b, and for $\mathcal{R}$ in 12.7c.

while 0 and 1 are better represented. The outcomes of the trait $\mathcal{R}$ -just as in $\mathcal{B}$ - reveal a large overlap between labels 0 and 1, and between 2 and 3.

Regarding the outcomes of Table 12.12, the $\mathcal{G}$ trait presents the best results as expressed by RMSE and MAE (0.69 and 0.51 respectively). Large errors are achieved with traits $\mathcal{R}$ and $\mathcal{B}$ (RMSE =0.76 and RMSE =0.72 respectively), which is in line to what is observed in the pdf curves, where a significant overlap occurs between labels 2-3. An analysis of the RMSE per label on each of the traits, reveals a pattern of large errors in the extremes. This might be due to the *more extended support* of labels 0 and 3 (3 units) compared to 1 and 2 (only 2 units), which is translated in the *possibility* of having larger errors that are more negatively penalised by RMSE .

12.2.2 Trial A2_DN

130 registers of the DN corpus (see details of the database in section 10.1.8) are employed for assessment. The results of the ordinal classification procedure are in Table 12.13. Similarly, the confusion matrices of the classification method providing the best results are presented in Figure 12.8.

Table 12.13. *Trial A2_DN*: best results obtained using the ordinal classification techniques and the DN partition.

	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
	f_β	MAE	AMAE	f_β	MAE	AMAE	f_β	MAE	AMAE
POM	0.51	0.45	0.48	0.51	0.51	0.51	0.40	0.83	0.73
ELMOP	0.53	0.43	0.47	0.48	0.57	0.58	0.39	0.84	0.77

The outcomes indicate an acceptable performance for $\mathcal{G}$ and $\mathcal{B}$ traits, and a decreased efficiency for $\mathcal{R}$. The POM technique tends to outperform ELMOP but in the $\mathcal{G}$ case. While the confusion matrices for the $\mathcal{G}$ trait suggest that most of the errors occur in the neighbour label, errors are more widespread when considering $\mathcal{B}$ and $\mathcal{R}$. The label 1 is the one with the worst results among trials, with SE ranging from 15.4% for $\mathcal{R}$, to 18.8% for $\mathcal{B}$. The label 0 is well modelled, achieving SE of 69.8% for $\mathcal{G}$, of 74.6% for $\mathcal{B}$ and of 60% for $\mathcal{R}$. Similarly the label 3 achieves SE ranging from 77.8% for $\mathcal{R}$ to 100% for $\mathcal{B}$. As in *trial A2_SVD*, the graphic that arises from plotting the weighted- AMAE for a range of weights is presented in Figure 12.9. In this case, outcomes suggest that accepting errors in the diagonals decreases AMAE to about 0.05 for $\mathcal{G}$, to 0.16 for $\mathcal{B}$ and to 0.40 $\mathcal{R}$.

To provide further evidence of the performance of the system, error plots after having modelled a Gaussian regressor are introduced in Figure 12.10. In addition, Table 12.14 presents the RMSE and MAE values that are obtained using the DN partition.

The outcomes of the Gaussian regressor indicate that the system predicts successfully $\mathcal{G}$ and $\mathcal{B}$ (RMSE =0.69 in both cases). However, results for $\mathcal{R}$ are not nearly as good (RMSE =0.96). As suggested by the RMSE calculated per label, the largest errors occur with the labels 0 and 3. By contrast, label 2 has the lowest error rates among all traits. Interestingly, there are prominent

(a) Confusion matrix for the $\mathcal{G}$ trait using ELMOP. (b) Confusion matrix for the $\mathcal{B}$ trait using POM. (c) Confusion matrix for the $\mathcal{R}$ trait using POM.

Figure 12.8. *Trial A2_DN*: confusion matrices for the ordinal classification (right) and their simulated counterparts (left) using the DN partition. Results for $\mathcal{G}$ are in Figure 12.8a, for $\mathcal{B}$ in 12.8b, and for $\mathcal{R}$ in 12.8c. The values next to the matrix indicate the SP and SE of the system. Each cell contains the number of incidences on top, and the percentage of occurrence at the bottom.

peaks using each feature suggesting certain consistency in the predicted label, however, label 3 tended to be centred towards 2-2.5 instead of getting closer to 3. It is worth noting that unlike the previous trial where at least 50 labels of each type (0, 1, 2, 3) are ensured, in this particular there are some cases with no more than 9 observations. The worst performance is provided by $\mathcal{R}$ where a large overlap is observed between labels 0, 1 and 2.

Figure 12.9. *Trial A2_DN*: weighted-AMAE for weights varied between 0 and 1. Results for the *DN* database.

Table 12.14. *Trial A2_DN*: error measures for the Gaussian regressor. Figure (a) presents the global values of RMSE and MAE, whereas (b) presents the RMSE per label within each trait.

	RMSE	MAE		0	1	2	3
$\mathcal{G}$	0.69	0.55	$\mathcal{G}$	0.89	0.59	0.51	0.62
$\mathcal{B}$	0.69	0.50	$\mathcal{B}$	0.80	0.63	0.45	0.75
$\mathcal{R}$	0.96	0.76	$\mathcal{R}$	1.14	0.82	0.50	1.01

(a) RMSE & MAE per class. (b) RMSE calculated for each trait and label.

12.2.3 *Trial A2_ATIC*

The results of the ordinal classification procedure using the *ATIC* dataset (see section 10.1.9) are introduced in Table 12.15. Similarly, the confusion matrices of the regression method providing the best results are in Figure 12.11.

Table 12.15. *Trial A2_ATIC*: best results obtained using the ordinal classification techniques and the *ATIC* partition.

	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
	f_β	MAE	AMAE	f_β	MAE	AMAE	f_β	MAE	AMAE
POM	0.42	0.67	0.69	0.52	0.51	0.58	0.45	0.69	0.88
ELMOP	0.40	0.72	0.74	0.41	0.67	0.74	0.43	0.88	0.89

Results indicate that *POM* techniques outperform *ELMOP*. In general the error rates in terms of AMAE are larger than in the two previous trials, with $\mathcal{B}$ providing the best results (0.58), and $\mathcal{R}$ the worst (0.88). This is also observed in the confusion matrices by means of a larger widespread of errors than in previous cases. Despite that, most of the errors are in the immediate neighbour and thus a certain tendency towards the diagonal is evidenced. To provide more hindsight about this, the graphic that arises from plotting the weighted-AMAE for a range of weights, and for the proposed and the artificial system, is presented in Figure 12.12. As observed, by allowing errors in the neighbourhood of the diagonal decreases AMAE to about 0.25

(a) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{G}$ trait.

(b) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{B}$ trait.

(c) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{R}$ trait.

Figure 12.10. *Trial A2_DN*: raster plots and pdf modelling the differences between the target label and the label predicted by the system. The deviations are normalised around the target label. Results for $\mathcal{G}$ are in Figure 12.10a, for $\mathcal{B}$ in 12.10b, and for $\mathcal{R}$ in 12.10c.

for $\mathcal{G}$, to 0.20 for $\mathcal{B}$ and to 0.60 for $\mathcal{R}$. In this case, large errors are still encountered for $\mathcal{R}$, but a good performance is obtained for $\mathcal{G}$ and specially $\mathcal{B}$, that even outperforms the optimistic artificial system.

(a) Confusion matrix for the $\mathcal{Q}$ trait using POM. (b) Confusion matrix for the $\mathcal{B}$ trait using POM. (c) Confusion matrix for the $\mathcal{R}$ trait using POM.

Figure 12.11. *Trial A2_ATIC*: confusion matrices for the ordinal classification (right) and their simulated counterparts (left) using the *ATIC* partition. Results for $\mathcal{Q}$ are in Figure 12.11a, for $\mathcal{B}$ in 12.11b, and for $\mathcal{R}$ in 12.11c. The values next to the matrix indicate the *SP* and *SE* of the system. Each cell contains the number of incidences on top, and the percentage of occurrence at the bottom.

The raster plots and pdf after having modelled using the Gaussian regressor are depicted in Figure 12.13, whereas Table 12.16 introduces the *RMSE* and *MAE* values that are obtained using the *ATIC* partition.

Like in the ordinal regression case, the error rates provided by the Gaussian regressor are larger than in previous trials, having *RMSE* values ranging from 0.62 for $\mathcal{B}$ to 0.95 for $\mathcal{R}$. Large errors are obtained for the label 3 in all

Figure 12.12. *Trial A2_ATIC*: weighted-AMAE for weights varied between 0 and 1. Results for the *ATIC* database.

Table 12.16. *Trial A2_ATIC*: error measures for the Gaussian regressor. Figure (a) presents the global values of RMSE and MAE, whereas (b) presents the RMSE per label within each trait.

	RMSE	MAE		0	1	2	3
$\mathcal{G}$	0.72	0.58	$\mathcal{G}$	0.53	0.62	0.73	1.00
$\mathcal{B}$	0.62	0.45	$\mathcal{B}$	0.37	0.42	0.76	1.02
$\mathcal{R}$	0.95	0.70	$\mathcal{R}$	0.96	0.42	0.94	1.42

(a) RMSE & MAE per class. (b) RMSE calculated for each trait and label.

cases as depicted in Table 12.16. The raster plots and pdf evidence how the label 3 is not fully represented, having a peak centred around 2.

12.2.4 Discussions on sub-test A2

The current sub-test has extended the findings of *sub-test A1*, by providing an examination about the most consistent features among all tested sets and databases. According to three filter feature selection algorithms, the best results, for all traits, are achieved by means of CHNR, GNE and RALA. When projecting these features in a 3-dimensional space, it can be observed that most of the overlap occurs among labels 0, 1 and 2, which contrasts with the good representation provided -in general- by label 3.

Trial A2_SVD constitutes a first approach to evaluate the system performance. It has to be noted, though, that using recordings for assessment which have also been used for modelling (although after being heavily processed and concatenated with data of other datasets), might inject information into the system about the experimental conditions followed during the recording of the speech samples, and therefore might bias results. Despite this shortcoming, it provides information about the system performance under "favourable" conditions. The results of the trial indicate that the ordinal classification works well on predicting the traits, specially $\mathcal{G}$ and $\mathcal{B}$. Thus, $\mathcal{G}$ achieves a MAE of 0.48, $\mathcal{B}$ of 0.56 and $\mathcal{R}$ of 0.64, indicating that on average labels deviate about half a unit from the perceptual label provided by the

(a) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{Q}$ trait.

(b) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{B}$ trait.

(c) Raster plot and pdf of deviations from labels for the Gaussian regression trial considering the $\mathcal{R}$ trait.

Figure 12.13. *Trial A2_ATIC*: raster plots and pdf modelling the differences between the target label and the label predicted by the system. The deviations are normalised around the target label. Results for $\mathcal{Q}$ are in Figure 12.13a, for $\mathcal{B}$ in 12.13b, and for $\mathcal{R}$ in 12.13c.

evaluator. AMAE is a measure that penalises errors in the ordering, in such a manner that the farther the predicted label from its target, the larger the AMAE . The error rates given by AMAE are in the same order of magnitude than those of MAE (0.50 for $\mathcal{Q}$, 0.54 for $\mathcal{B}$ and 0.63 for $\mathcal{R}$), suggesting that, in

general, errors should be located in the neighbour labels. This can be verified by examining the confusion matrices presented in Figure 12.5, where most of the errors are concentrated around the diagonal. This is further demonstrated in the weighted-AMAE graphics, which indicate a drastic drop in the AMAE values when errors in the neighbourhood of the diagonals are treated as positive instances. In particular, setting the weight to zero decrements errors to values ranging from 0.15-0.20 in the SVD corpus, to 0.05-0.40 in DN, whereas in ATIC they vary in the range 0.20 to 0.60. In contrast to *trial A2_SVD*; *trial A2_DN* and *trial A2_ATIC* are interesting experiment which utilise external datasets for assessment purposes, avoiding the biasing problems and challenging the generalist capabilities of the proposed system in a cross-database setting. Particularly for *trial A2_DN*, the outcomes suggest some consistency compared to the results of *trial A2_SVD*. As previously, the performance of the ordinal classifier is much better for $\mathcal{G}$ (AMAE=0.47) and $\mathcal{B}$ (AMAE=0.51), than for $\mathcal{R}$ (AMAE=0.73). Similarly, the confusion matrices indicate that most of the errors are concentrated around the diagonal, but with more dispersion than in *trial A2_SVD*. Finally, *trial A2_ATIC* serves as a second database that is employed for testing out the performance that the system achieves. Outcomes follow the same line of the experiments using the DN and SVD database, i. e., the ordinal classification achieve better results with $\mathcal{G}$ and $\mathcal{B}$. Notwithstanding, this trial presents the largest dispersion in results as observed from the confusion matrices.

The Gaussian regressor agrees in most of cases to the outcomes provided by the ordinal classifier. Particularly for *trial A2_SVD*, MAE achieves values of $\mathcal{G}$ = 0.51, $\mathcal{B}$ = 0.57 and $\mathcal{R}$ = 0.58, which are close to the results reported using the ordinal classifier. RMSE -which penalises large deviations from the target- indicates an affected performance specially for the $\mathcal{R}$ trait. For *trial A2_DN*, the behaviour of the Gaussian regressor follows the same line of the ordinal classifier with $\mathcal{G}$ and $\mathcal{B}$ presenting better performance than $\mathcal{R}$ (MAE for $\mathcal{G}$ of 0.55, $\mathcal{B}$ of 0.50 and $\mathcal{R}$ of 0.76). Likewise, for *trial A2_ATIC*, $\mathcal{R}$ is reported once again as the trait with the lowest performance (MAE=0.70 and RMSE=0.95), whereas in this case $\mathcal{B}$ provides the best efficiency (MAE=0.45 and RMSE=0.63). This trial present some of the largest dispersion in results as evidenced by the pdf and raster plots. Finally, it is also worth noting, that when analysing the distribution of errors in the Gaussian regressor (Figures 12.7 and 12.10) two major clusters are observed: one related to errors between labels 0 and 1, and another between 1 and 2. It is difficult to ascertain the reason of this behaviour, but a hypothesis stands on the criteria followed by the evaluator to grade each one of the traits. In her words, "*the labels 0 and 2 are straightforward to assign, as the first indicates the absence of pathology, while the latter ensures the presence of disorders. In the same manner, label 3 contains the most extreme cases of pathology and accounts for those conditions where the voice will unlikely get worse*". This statement indicates that there is a natural frontier between the actual existence of the disorder, and its absence or low likelihood: label 2. Indeed, label 1 is a value in-between 0 (total absence) and 2 (confirmed pathology) which embodies those scenarios where is difficult to provide a decision whether there exist the presence of a pathology or not.

A common pattern of all the trials is the good performance of $\mathcal{G}$ and $\mathcal{B}$, compared to $\mathcal{R}$. One hypothesis explaining this behaviour might be in the use of $\mathcal{R}$ ALA, $\mathcal{G}$ NE and $\mathcal{C}$ HN $\mathcal{R}$ for modelling the system, as these features are designed to capture additive noise characteristics, and as such, should provide more information about breathiness than roughness. The inability of the filter algorithms to include multidimensional features (such as $\mathcal{P}$ LP and $\mathcal{M}$ FCC), which might simultaneously add information to the system about roughness and breathiness, might have penalised the performance of the $\mathcal{R}$ trait. Another hypothesis could be in a certain bias of the evaluator towards taking a decision regarding roughness, or the presence of pathologies that affect the $\mathcal{B}$ and $\mathcal{R}$ perception simultaneously.

12.3 SUB-TEST A3: INFLUENCE OF SEX IN AUTOMATIC ASSESSMENT TASKS.

Sub-test A3 examines the effects of accounting for the sex of the speaker in the design of automatic assessment systems. The methodological stages that are followed are the same of the previous sub-test and have already presented in section 10.2.2.2. The only difference with the previous sub-test is in having designed two sex-dependent systems that are built using the evaluations of female and male speakers separately. These results are then combined, providing a single decision about performance. In a similar fashion to *sub-test A2*, three trials are performed using a partition of $\mathcal{S}$ VD, $\mathcal{A}$ TIC and the $\mathcal{D}$ N corpora.

12.3.1 Trial A3_SVD

A partition of 121 registers of the $\mathcal{S}$ VD dataset is employed for assessment purposes as described in subsection 10.2.2.2. $\mathcal{P}$ OM and $\mathcal{E}$ LMOP ordinal classifiers are trained for testing out the ability of the features selected in *sub-test A2* to predict $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$. The combined outcomes of the sex-dependent systems are introduced in Table 12.17.

Table 12.17. *Trial A3_SVD*: best results of the sex-dependent system obtained using the ordinal classification techniques and the $\mathcal{S}$ VD testing partition.

	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
	f_β	MAE	AMAE	f_β	MAE	AMAE	f_β	MAE	AMAE
$\mathcal{P}$ OM	0.53	0.58	0.58	0.45	0.67	0.64	0.42	0.72	0.71
$\mathcal{E}$ LMOP	0.53	0.59	0.57	0.44	0.67	0.64	0.44	0.73	0.72

The best results are obtained when using the $\mathcal{E}$ LMOP classifier for the $\mathcal{G}$ and $\mathcal{B}$ traits, whereas $\mathcal{P}$ OM performs better when using $\mathcal{R}$. The outcomes indicate the same behaviour of previous sub-tests, i. e., $\mathcal{G}$ presents the smallest error rates, whereas $\mathcal{R}$ has the largest ones.

12.3.2 Trial A3_DN

Table 12.18 introduces the performance metrics of the sex-dependent systems trained with the features selected in *sub-test A2*, POM and ELMOP ordinal classifiers, and using registers of the DN database.

Table 12.18. Trial A3_DN: best results of the sex-dependent system obtained using the ordinal classification techniques and the DN testing partition.

	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
	f_β	MAE	AMAE	f_β	MAE	AMAE	f_β	MAE	AMAE
POM	0.50	0.47	0.50	0.50	0.51	0.52	0.40	0.81	0.74
ELMOP	0.51	0.50	0.50	0.48	0.52	0.61	0.41	0.75	0.71

As previously, $\mathcal{G}$ is the system presenting the best performance whereas $\mathcal{R}$ has the worst. POM outperforms ELMOP for the $\mathcal{B}$ trait, whereas the contrary occurs for $\mathcal{R}$. Results are comparable for both classifiers when $\mathcal{G}$ is considered.

12.3.3 Trial A3_ATIC

For assessment purposes a partition of 121 registers of the ATIC dataset is employed as described in subsection 10.2.2.2. POM and ELMOP ordinal classifiers are trained for testing out the ability of the features selected in *sub-test A2* to predict $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$. The combined outcomes of the sex-dependent systems are introduced in Table 12.19.

Table 12.19. Trial A3_ATIC: best results of the sex-dependent system obtained using the ordinal classification techniques and the ATIC testing partition.

	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
	f_β	MAE	AMAE	f_β	MAE	AMAE	f_β	MAE	AMAE
POM	0.48	0.61	0.64	0.54	0.48	0.55	0.47	0.67	0.85
ELMOP	0.39	0.77	0.78	0.43	0.67	0.73	0.39	0.65	0.91

In this particular case, $\mathcal{B}$ presents the lowest error rates, in contraposition to the previous trials on which $\mathcal{G}$ had them. Despite that, $\mathcal{R}$ is still presenting the worst results among traits. Results also indicate that POM is the classifier having the best performance.

12.3.4 Discussions on sub-test A3

The current sub-test has been performed to analyse the effects of the speaker's sex in the design of hierarchical assessment systems. This follows the idea presented in section 10.2.1.1 where extralinguistic traits are employed to simplify the decisions of an automatic detection system. The methodology that has been used is similar to the one introduced in *sub-test A2* but considering a male-only and a female-only assessment systems, which are then validated in three databases (partition of SVD, ATIC and DN)

Table 12.20. *Sub-test A3*: comparison of the results of the sex-dependent (S.D.) and sex-independent (S.I.) assessment systems in terms of AMAE.

	$\mathcal{G}$		$\mathcal{B}$		$\mathcal{R}$	
	S.I.	S.D.	S.I.	S.D.	S.I.	S.D.
SVD	0.50	0.57	0.54	0.64	0.63	0.71
DN	0.47	0.50	0.51	0.52	0.73	0.71
ATIC	0.69	0.64	0.58	0.55	0.88	0.85

Results are then combined and performance metrics are extracted. For the sake of comparison, the results -in terms of AMAE- of the sex-independent system presented in *sub-test A2*, and the hierarchical sex-dependent systems of the current sub-test are introduced in Table 12.20.

On contrary to the outcomes obtained in the detection experiments, it has been found that in most of the cases the sex-independent system provides better results than the sex-dependent one. Indeed, this behaviour is observed in all the databases but ATIC, where the performance favours the sex-dependent case but where the differences are not significant. The largest performance divergence between approaches arises in the SVD database, where the sex-independent system largely outperforms the sex-dependent case, e.g., for $\mathcal{B}$ AMAE is of 0.54 whereas for the latter it is of 0.64. Due to this results and due to the need of a more complex design methodology, the usage of sex-dependent systems are disregarded from further testing.

12.4 SUB-TEST A4: CLINICAL VALIDATION OF GAUSSIAN REGRESSION SYSTEMS.

The clinical validation of the continuous systems presented in *sub-test A2* is to be performed. The methodology that is followed has been presented in section 10.2.2.4. The speech evaluator is asked to perform two tasks. First, to provide blind assessments to the ATIC, SVD and DN datasets, but redefining the GRBAS scale in a continuum. Second, the results of the automatic systems in *trial A2_SVD*, *trial A2_DN* and *trial A2_ATIC* are given to the speech therapist to assess the correctness of the predictions. The evaluator receives the computer application portrayed in Figure 10.17, having the slider of each trait set to the value predicted by the automatic system. Registers are then re-evaluated, in such a manner that if the evaluator disagrees with any of the labels it is asked to adjust it. After this procedure is carried out, error metrics are computed between the predicted labels and the new evaluations corrected by the evaluator. Results in terms of MAE and RMSE for this second task are presented in Table 12.21, for the SVD, ATIC and DN corpora.

To provide more hindsight on the distribution of errors the deviation between predicted and corrected labels is presented in Figure 12.14 for the SVD partition, in 12.15 for DN, and in Figure 12.16 for ATIC.

Table 12.21. *Sub-test A4*: error measures between the labels predicted by the automatic system and the corrections made by the speech therapist.

	$\mathcal{G}$		$\mathcal{B}$		$\mathcal{R}$	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
SVD	0.48	0.32	0.59	0.47	0.67	0.49
DN	0.41	0.29	0.50	0.38	0.69	0.48
ATIC	0.46	0.27	0.47	0.26	0.50	0.30

Figure 12.14. *Sub-test A4*: raster plots and pdf of the deviation between the evaluations predicted by the system in *sub-test A2* and the corrected labels provided by the speech therapist for the SVD database.

Figure 12.15. *Sub-test A4*: raster plots and pdf of the deviation between the evaluations predicted by the system in *sub-test A2* and the corrected labels provided by the speech therapist for the DN database.

12.4.1 Discussions on sub-test A4

Before the actual discussion of the results obtained in the current sub-test, an analysis about the consistency that the speech therapist provides in the evaluations is firstly considered. This constitutes an exploratory analysis which indicates up to what extent it is possible to emulate the perceptual capabilities of the expert. Indeed, if the evaluator presents a low consistency between evaluations made on the same voice material, the system will be

Figure 12.16. *Sub-test A4*: raster plots and pdf of the deviation between the evaluations predicted by the system in *sub-test A2* and the corrected labels provided by the speech therapist for the ATIC database.

probably conditioned by the introduction of variability due to this lack of coherence. To perform this analysis, all the available assessments provided by the speech therapist in the different databases employed in this thesis are considered. However, and as referred in Table 12.22, only those datasets containing more than one assessment are used as it is necessary to perform comparisons of evaluations using the same acoustic material.

Table 12.22. Databases used for evaluating the consistency of the expert. The number and type of assessments for each database is presented.

Database	Number of evaluations	
	Discrete	Continuous
SVD	1	2
DN	1	2
ATIC*	1	1

*ATIC has an extra assessment obtained in consensus by three experts as indicated in section 10.1.9.

Several tests are performed in different scenarios. First, the continuous evaluations provided by the expert during the first task of *sub-test A4*, using the SVD partition and the DN corpus, are measured by means of the generalisability coefficient ($\bar{U}$) and the distance correlation ($\rho_{\mathcal{R}}$). The results of this analysis are presented in Table 12.23.

Table 12.23. *Sub-test A4*: consistency between continuous assessments of the reduced GRBAS scale for the SVD partition and DN database.

	$\mathcal{G}$		$\mathcal{B}$		$\mathcal{R}$	
	$\bar{U}$	$\rho_{\mathcal{R}}$	$\bar{U}$	$\rho_{\mathcal{R}}$	$\bar{U}$	$\rho_{\mathcal{R}}$
SVD	0.93	0.85	0.89	0.78	0.91	0.81
DN	0.94	0.87	0.94	0.87	0.85	0.73

As observed, there is a large consistency among all traits and databases, being specially large for $\mathcal{G}$ where $\bar{U} = 0.94$, but with remarkably high values for $\mathcal{B}$ ($\bar{U}$ up to 0.94) and $\mathcal{R}$ ($\bar{U}$ up to 0.91) too. This suggests that the speech therapist provides accurate descriptions of $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$, even after having considered the grading process as a continuum. To compare the consistency of these continuous gradings with the discrete labels, three corpora are examined: the partition of **SVD**, **DN** and **ATIC**. The results of consistency between the discrete and continuous evaluations are introduced in Table 12.24.

Table 12.24. *Sub-test A4*: consistency between assessments using the discrete and continue gradings of the reduced **GRBAS** scale for the partition of **SVD**, **DN** and **ATIC**. *D.* stands for discrete, *C1.* for the first continuous evaluation and *C2.* for the second.

	SVD			DN			ATIC	
	$\rho_{\mathcal{R}}$			$\rho_{\mathcal{R}}$			$\rho_{\mathcal{R}}$	
	$\bar{U}$	<i>D.-C1.</i>	<i>D.-C2.</i>	$\bar{U}$	<i>D.-C1.</i>	<i>D.-C2.</i>	$\bar{U}$	<i>D.-C1.</i>
$\mathcal{G}$	0.94	0.84	0.83	0.93	0.73	0.78	0.92	0.81
$\mathcal{B}$	0.86	0.65	0.70	0.93	0.73	0.75	0.84	0.66
$\mathcal{R}$	0.89	0.73	0.72	0.85	0.60	0.61	0.86	0.70

It might be expected that values of consistency and correlation would decrease with respect to the previous case, due to having compared a continuous and a discrete scale that only permits a limited number of quantisation steps (4 values). This indeed, might induce subsequently larger errors that produce a consequent decrease in consistency. Notwithstanding, for the $\mathcal{G}$ trait, the generalisability coefficient is similar when the **SVD** dataset is analysed. Furthermore, $\rho_{\mathcal{R}}$ maintains the same levels of correlations indicating a surprisingly large ability of the evaluator to ascertain evaluations for this particular trait. When the remaining traits are analysed the expected decrease is observed. Despite that, it can be inferred that in general the evaluator is consistent in her assessments. So far these tests have provided an insight about the *accuracy* of the evaluator but no clue about its *precision* has been provided. To this end, and in an attempt to correlate the results of the discrete evaluations provided by the speech therapist to those formed by a consensus of three speech pathologists, the **ATIC** database is employed. Results of this consistency analysis are presented in Table 12.25.

Table 12.25. *Sub-test A4*: consistency between the evaluations provided by the speech therapist and three experts in consensus for the **ATIC** corpus.

$\mathcal{G}$		$\mathcal{B}$		$\mathcal{R}$	
$\bar{U}$	$\rho_{\mathcal{R}}$	$\bar{U}$	$\rho_{\mathcal{R}}$	$\bar{U}$	$\rho_{\mathcal{R}}$
0.67	0.48	0.71	0.62	0.52	0.28

A remarkably low consistency between the evaluations of the speech therapist and the consensus of evaluators is observed, having the largest values in $\mathcal{B}$ ($\bar{U} = 0.71$) and the lowest in $\mathcal{R}$ ($\bar{U} = 0.52$). In all cases these results

reveal a lack of agreement. In terms of $\rho_{\mathcal{R}}$ these values vary between 0.28 to 0.48, being indicators of poor correlation among assessments in this scenario. Interestingly the worst agreement -by a large amount- is in the $\mathcal{R}$ trait, which is line to what is observed in the automatic system of *sub-test* A2. Nothing else can be concluded from this analysis except that the criteria used by our speech therapist and the consensus of experts differ.

Now it is possible to discuss the results of the trials presented in *sub-test* A2, having confidence in the ability of the evaluator to provide consistent and repeatable assessment results. In this respect and as indicated in Table 12.21, the errors of the automatic system tend to be small in most of the cases. Particularly for the SVD partition, and even though the values of $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$ are in all cases satisfactory, the best performance is obtained using the $\mathcal{G}$ trait, and the worst is observed for the $\mathcal{R}$ trait. Notably, RMSE varies between 0.48 (for $\mathcal{G}$) to 0.67 (for $\mathcal{R}$), whereas MAE is in the range of 0.32 to 0.49. The good behaviour of $\mathcal{G}$ is also observed in Figure 12.14, where the raster plot is centred around 0 and a large peak is obtained in the pdf plot. By contrast, $\mathcal{R}$ and $\mathcal{B}$ present a flatter and more widespread pdf. Regarding the DN partition, RMSE varies between 0.41 for $\mathcal{G}$, to 0.69 for $\mathcal{R}$; whereas MAE ranges from 0.29 for $\mathcal{B}$ to 0.48 for $\mathcal{R}$. Similarly to the previous case, errors tend to be larger for $\mathcal{R}$ than for the other two traits, whereas $\mathcal{G}$ has the best general performance. This behaviour is also observed in Figure 12.15, where $\mathcal{G}$ presents an uni-modal peak centred around zero. In comparison, the distribution of $\mathcal{B}$ and $\mathcal{R}$ is more widespread. It is noticeable that in the raster plot of $\mathcal{R}$, errors larger than 2 units are produced. Finally for the ATIC database, outcomes indicate that the system tends to behave similarly to the other datasets for $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$. Indeed, RMSE varies between 0.46 for $\mathcal{G}$ and 0.50 for $\mathcal{R}$; whereas MAE ranges from 0.26 for $\mathcal{B}$ to 0.30 for $\mathcal{R}$. The discrepancies between RMSE and MAE values indicate that although errors can be considered -generally- small (labels deviate by about 0.3 units), there are some cases where large mistakes are committed, thus increasing the RMSE values. This can be observed in the raster plot of the $\mathcal{B}$ trait where errors as big as 2 units are present. Despite that, the system behaves -in general terms- satisfactorily, as the raster plots and the pdf of deviations demonstrate (Figure 12.16), having most of the errors centred around 0.

12.5 GENERAL DISCUSSION

The current section has presented experiments related to the automatic assessment of voice pathologies. The goal has been to generalise results by using several databases in conjunction to diverse sets of features, and to train an automatic system capable of modelling the perceptual capabilities of a human evaluator. The outcomes indicate that it is possible to design, with a certain degree of success, systems that correctly predict the $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$ traits according to the perceptual criteria given by a human evaluator. It has been found that the most informative and consistent features amongst the trials are CHNR, GNE and RALA, for all $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$; hence resembling the outcomes of *sub-test* D3. This is reasonable considering that there is a

close relationship between pathological states and perceptual evaluations of voice quality. Indeed, $\mathcal{G}$ might be regarded as a trait that evaluates the general state of a certain speaker i. e., it characterises the degree of hoarseness present in voice, being a good descriptor of pathology. The fact that the same features are regarded informative in conjunction for $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$ might have been the result of particularities of the evaluations or due to the presence of pathologies affecting the perceptions of $\mathcal{B}$ and $\mathcal{R}$ simultaneously. However it is difficult to conclude about this particular and further testing is required.

Ordinal classification errors vary among trials, with $\mathcal{AMAE}$ ranging from 0.47 to 0.69 for $\mathcal{G}$, from 0.51 to 0.58 for $\mathcal{B}$, and from 0.63 to 0.88 for $\mathcal{R}$. These values are of the same order of magnitude as those of $\mathcal{MAE}$, for which it is possible to infer that -on average- the deviations between the predicted and the actual label are less than one label apart from each other (or half a label in the case of $\mathcal{G}$ and $\mathcal{B}$). When a Gaussian regressor is considered to account for the continuous nature of the $\mathcal{GRBAS}$ scale, errors have been found to be approximately in the same orders of magnitude of the ordinal classification. Similarly, it has been found that the largest errors are found when $\mathcal{R}$ is analysed. This phenomenon might be expected as $\mathcal{GNE}$ and $\mathcal{CHNR}$ are characteristics intended to measure additive noise that is consequence of turbulence, i. e., they are more interrelated to perceptual variations due to breathiness than to roughness. In a similar fashion, $\mathcal{RALA}$ is a parameter that measures the dispersion of energies in the $\mathcal{MS}$, and as such it would be reasonable to consider it an estimator of "harmonicity", and hence, of additive noise components. Despite all three features are intended as estimators of turbulence, it is important to remark that phenomena such as jitter and shimmer (descriptors of modulation noise), also induce a loss in periodicity that affects the harmonic components of speech. As a result, it is expected that these three measures not only characterise additive noise but also account indirectly for modulation components as harmonic behaviour is a central part in their calculation.

One important consideration regarding the perceptual evaluation using the $\mathcal{GRBAS}$ scale, is in the subjectivity of the evaluations. Indeed, several factors have been known to affect assessments, including the evaluator's experience, the perceptual rating scale that is used, the speech task, fatigue, sensitivity of the listener to the vocal trait that is analysed, etc. [336]. Even though the databases have been evaluated by the same evaluator, it is possible that the ratings might have been affected by all these subjective factors as they are intrinsic to the evaluation process itself. In addition, it is worth noting that the evaluations of the $\mathcal{SVD}$, $\mathcal{HUPA}$ and $\mathcal{GMar}$ databases have been made at different moments in time -even with some years of difference among them- for which all these above described effects might have been accentuated. Despite these shortcomings, the evaluator has shown an impressive intra-evaluation consistency, which demonstrates its knowledge of a well-defined evaluation criterion that is accurate and allows repeatability among evaluations. Notably, the large differences between the assessment provided by the committee in $\mathcal{ATIC}$ and the evaluation of the speech therapist, highlights the subjectivity of the assessment problem, demonstrating the difficulties

of achieving good inter-evaluator agreements. Certainly, as discussed before, we hypothesise that one possible reason of the decreased performance obtained with the $\mathcal{R}$ trait might have been attributable to the perceptual sensitivity of the evaluator to this particular trait.

Outcomes of *sub-test A4* suggest that redefining the GRBAS scale into a continuum has not gravely impacted the assessment task. Indeed, it has been found a generalisability coefficient that is approximately of the same order of magnitude, when comparing the consistency of the continuous evaluations (Table 12.23) to the continuous and discrete gradings (Table 12.24). In particular, and for the first case, the worst results in terms of $\mathcal{U}$ range between 0.85 for $\mathcal{R}$ to 0.93 for $\mathcal{G}$, whereas for the latter it ranges between 0.84 for $\mathcal{R}$ to 0.85 for $\mathcal{B}$. In general these values indicate that the assessments made by the speech therapist present a large intra-evaluator agreement.

The regression procedure has been shown to be effective in providing reliable assessments which are in the same order of magnitude to the results given by the ordinal classifier. In addition, it has been shown to be reliable emulating the perceptual capabilities of the evaluator as ascertained during the clinical evaluation of results, where an agreement between the decisions taken by the expert and the ones provided by the automatic system has been demonstrated. Particularly, the largest errors between the predicted labels and the relabelled system revolves, in terms of MAE, around 0.27 for $\mathcal{G}$, 0.26 for $\mathcal{B}$ and 0.30 for $\mathcal{R}$, occurring in all cases in the ATIC database. This is specially interesting, as the ATIC database presented in *trial A2_DN* some of the more disperse results when the Gaussian regression was considered.

Literature reports other types of systems where the automatic assessment of voice pathologies has been studied. However, it is not possible to provide direct comparisons to those works as the methodologies that are followed differ drastically in terms of the techniques that are used and of the databases that have been considered. For instance, in [125], continuous speech and sustained vowels have been used to construct models correlating with overall voice quality. In [337], a system based on ANN, spectral/cepstral and perturbation features has been used to provide assessments of laryngeal cancer to a series of patients evaluated according to a 7-rating scale of severity, obtaining an efficiency of 92% in the assessment task. However, no cross-validation of the results is employed, and as stated by the authors “...as a result of the relatively small dataset, there were different numbers of patients in each class. As it is desirable to have equal numbers in each class to train an ANN adequately, additional frames were taken from some patients in classes with the fewest patients”. This indicates that it is possible that the frames of certain patients are used either for training and for testing the system, for which the results might have been biased. In [133], a system classifying between normophonic, moderately hoarse or severely hoarse voice using signal-to-dysperiodicity ratios and SVM is presented. The methodology is tested on a Dutch corpus of connected speech for which efficiency is about 70%. In [338], MFCC are used to predict traits in the HUPA database via learning vector quantization. The reported ACC for this methodology is around 65% but not cross-validation has been used to assess results. Perceptual evaluations have also been used

as auxiliary material for systems detecting voice pathology in the form of a-priori or complementary information injected to the system [38, 339]. An interesting approach is introduced in [340], where 65 perturbation, spectral/-cepstral and complexity features are used predict $\mathcal{G}$ in a Chinese database evaluated according to the GRBAS scale. Feature extraction techniques, and SVM or ELM classifiers are used for providing decision, obtaining accuracies of around 80%. As it can be inferred all above-mentioned systems provide a pattern recognition approach that differs to the one presented in the present experiment, and highlight the novelty of the followed approach.

Finally, it is worth noting that the ordinal classifiers and the Gaussian regressor techniques that have been used in this thesis have been chosen due to their simplicity. However, other algorithms can be tested out in order to improve performance. In particular, POM is a classical and reliable linear technique that dates back to the 80's, but which is unable to handle nonlinearity in the feature space. Likewise, ELMOP has been found to be affected by small database size [311] -as it might have been the case when the per-file statistics has been used-, and which might be the reason explaining why it has been outperformed by the simpler POM algorithm. With this in mind, other types of techniques that should be considered in the future include Gaussian processes for ordinal classification, several modifications of the SVM algorithm to handle the ordinal classification problem, or other ANN-based approaches.

Part V

CONCLUSIONS AND FUTURE WORK

CONCLUSIONS AND FUTURE WORK

13.1 CONCLUSIONS

AVQA systems have been gaining a great interest due to their potential applications in clinical settings, where they might turn into valuable complementary tools for improving diagnosis and screening processes. In contrast to traditional procedures, these systems provide an objective assessment of the clinical state of patients, reducing the evaluation time and the cost of diagnosis and treatment. Furthermore, they avoid invasive procedures by employing signals which are easily recorded by inexpensive means. Nonetheless, due to the intrinsic variability of the speech and voice its performance can be compromised. Because of the importance of the topic, the aim of this thesis is in the design of AVQA systems based on speech technologies, capable of generalising decisions regarding the presence or the severity of voice pathologies. To this end, a variety of trials have been performed under different settings, aiming at evaluating the consistency of features, classifiers, methodologies and effect of the speech tasks in automatic pathology detection and assessment labours.

Results demonstrate that including extralinguistic information respecting the sex of the speaker enhances the performance of an automatic detector of pathologies. Nothing can be concluded about incorporating age information to the system as there is not enough evidence supporting its use, mostly due to the limitations of the available databases. In any case, the outcomes suggest that decomposing the pathology detection problem into sub-problems according to a certain extralinguistic criterion, decreases complexity and therefore improves efficiency of the systems.

It has been reported through the analysis of several sets of features and speech material of different types, that no single parameter is capable of completely characterising vocal pathology, and therefore multidimensional approaches are needed to enhance classification results. The most coherent features, as ascertained in several trials, are two estimators of perturbation noise and one descriptor of dispersion based on MS; namely, GNE, CHNR and RALA. Interestingly, the latter is a novel characteristic that has been recently introduced in [219, 220]. These three parameters are consistently better than the remaining features in both detection and assessment tasks, standing as reliable predictors of $\mathcal{G}$, $\mathcal{B}$, $\mathcal{R}$, and useful descriptors of dysphonic and normophonic behaviour. Even though these features are computed in different

domains, they have shown complementarity. This is evidenced by the outcomes that are obtained when these three features are used jointly. By contrast, characteristics extracted using a non-uniform embedding approach evidence a lack of consistency. Indeed, in some cases the non-uniform embedding method outperforms the classical uniform approach but there are many cases where the contrary occurs. Due to this fact and because of the increased computational burden of calculating lag vectors, this type of attractor reconstruction is not recommended -in its current form- for the design of AVQA systems. In addition to that, it has been found -through the comparison of regularity measures computed using the non-uniform and uniform approaches- that the parameters providing the best results in the characterisation of regularity are not necessarily those which are typically suggested in literature ($d_e=2$ and $\tau=1$). Instead, an optimisation should be followed to find the optimal parameters capable of characterising normophonic and dysphonic behaviour. Nonetheless, an in-depth study on this particular is still an open issue.

In reference to the types of speech tasks that have been evaluated, it has been discovered that methodologies based on the sustained vowel /a/ always achieve better results compared to using other types of sustained phonation. Notwithstanding, these results are outperformed when running speech is considered. Outcomes indicate that using a voiced detector to extract voicing fragments do not provide a performance enhancement compared to the parametrisation of the raw speech signal when spectral and cepstral features are used. It has also been found that when the acoustic material of different speech tasks is fused -via logistic regression- the performance of the system is enhanced even further. This fusion stage has demonstrated its utility to combine information of heterogeneous but complementary systems, into a single decision machine that behaves better than the individual systems.

UBM-based classifiers have moderately improved results for those cases where sustained vowels are used in conjunction with ancillary corpora based on normophonic phonation. By contrast, nothing can be concluded regarding the use of running speech and the UBM-based classifiers, as the phonetic mismatch between target and ancillary databases has presumably affected the resulting UBM and compensation models. Moreover, due to the unavailability of datasets containing dysphonic registers of running speech, an analysis on the effects of this type of acoustic material on the training process of compensation models remains an open issue. Despite that, it is safe to say that the results reported in this thesis involving the usage of UBM-based classifiers and running speech are not positive.

In general terms, it can be concluded that a methodology based in hierarchical detection, characterisation through a reduced set of consistent features and fusion of different speech tasks enhances performance. Indeed, a system trained with the SVD database and following the proposed approach achieves an ACC of 81%, which is one of the best results reported in literature for this corpus. In a cross-database scenario the AUC of the system

reaches values ranging from 0.75 to 0.94, demonstrating its robustness to handle mismatches between training and testing conditions.

With respect to the generalist assessment system, the most informative and consistent features are -as in the case of the detection system- [CHNR](#), [GNE](#) and [RALA](#). These three features are equally relevant to the characterisation of $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$. When this three features are plotted altogether, a very differentiated cluster is observed for the label 3 but with some interweaving for the remaining labels. In general, it has been found that the best results of the assessment system are, almost in all cases, provided by the $\mathcal{G}$ trait, whereas the worst are yielded by $\mathcal{R}$. This phenomenon has been hypothesized to occur due to the use of [GNE](#), [CHNR](#) and [RALA](#) which are estimators of turbulence noise, an as such are more interrelated to perceptual variations of breathiness than to roughness.

Results also indicate that the ordinal classifier has behaved satisfactorily in all the trials involving the [GRBAS](#) scale, providing consistent results which are in line to those provided by the Gaussian regressor. Because of that fact, using ordinal classification is recommended for the analysis of rating scales where the ordering of the assessment should be accounted. Similarly, it has been found that the distance correlation, $\rho_{\mathcal{R}}$, provides results which are consistent to other techniques used to analyse performance (such as the ordinal classifiers, feature ranking algorithms, etc.). Due to its consistency, simplicity, and its ability to characterise nonlinear relationships between variables (even in multidimensional scenarios), its usage is recommended in applications where the strength between variables should be gauged. Indeed, it can either replace or supplement other more classical correlation measures such as ρ , Spearman or Kendall's tau.

In most of the assessment experiments involving ordinal regression the errors occur in the immediacy of the target label. If the system is designed to accept committing this type of errors, a drastic drop in errors is generally obtained as ascertained by the weighted-[AMAE](#). Outcomes also indicate that is possible to extend the discrete [GRBAS](#) scale to a continuum, having found evidences about the consistency that an evaluator can achieve when using this redefined scale, and about the close relationship that the continuous labels presents to the standard discrete ratings. Finally, it can be concluded that it is possible to effectively emulate the perceptual capabilities of a human evaluator through techniques based on speech technologies. Indeed, the proposed system achieves errors -between the discrete labels given by the evaluator and the predictions of the systems- which are about half an unit, on average, when the $\mathcal{G}$ and $\mathcal{B}$ traits are considered. This system has also been proficient in generalising results as indicated by respectable [MAE](#) values in the cross-database experiments. In addition, the clinical validation of the system demonstrates the utility of the system in the prediction of $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{R}$. In particular for $\mathcal{G}$, [MAE](#) varied in the range 0.27 to 0.32. These values are lower than the results obtained with the proposed system, indicating that the expert considers that in most of cases the predictions given by the proposed system are in the range of what can be considered reasonable.

13.2 CONTRIBUTIONS

This thesis has presented a through study of diverse factors that might influence the performance of pathology and assessment systems. As a consequence of this work, the following original contributions can be listed:

- An analysis about the influence of extralinguistic factors such as age and sex has been performed, focusing on the impact in performance of hierarchical system guided by extralinguistic information.
- An analysis on the impact that several speech tasks produces in the performance of automatic detection systems has been carried out.
- A comparison on the effects of using diverse sets of features and acoustic material of different type for the detection of voice pathologies, has been performed.
- Classification systems based on UBM-classifiers, which are often used in speaker recognition applications, have been translated to the context of voice pathology detection. For experiments considering sustained phonation and running speech, several ancillary databases have been employed for training UBM and compensation models.
- A novel approach has been proposed to emulate the perceptual capabilities of a human evaluator, based on finding a reduced set of consistent characteristics through feature ranking algorithms. The reduced set of features is then used to build automatic systems aimed at generalising results by means of ordinal classification and Gaussian regression techniques.
- To our knowledge, this is the first time that ordinal classification has been employed in the context of perceptual evaluations of voice quality. These techniques have been shown to be more appropriate than classification, as they account for the ordinal character of rating scales as GRBAS. Likewise, descriptors such as $\rho_{\mathcal{R}}$ which measure multidimensional and nonlinear relationships are introduced to the AVQA field.
- The GRBAS has been redefined through the usage of techniques based on Gaussian regression. Results indicate that this method behaves favourably, having outcomes that are in line to the ones obtained with the classical discrete scale and ordinal classification.
- An analysis about the consistency of an evaluator providing continuous and discrete labels has been made, demonstrating the feasibility of extending the GRBAS scale to a continuum, and the close relationship that is obtainable when continuous and discrete ratings are considered.
- A series of cross-database experiments has been performed for either detection and assessment of voice pathologies. As a result, the performance of the proposed systems have been tested under unfavourable conditions of mismatch between training and testing instances.

13.3 FUTURE WORK

There are a lot of open issues that should be considered for future study:

- Further testing is required to extract evidences about the influence of the speaker's age in [AVQA](#) systems. This include an in-depth analysis to correlate chronological to physiological age, and to establish frontiers among age groups. Similarly, the study of paralinguistic information such as mood and accent (when using running speech) remains an open issue that should be studied. Likewise, novel approaches to include extralinguistic (or paralinguistic) information in the form of a-priori probabilities or in any other form should be readily explored.
- Since speech is a highly informative signal, a preprocessing stage that removes irrelevant components should be followed. This ensures that the characterisation stage is focused on the phenomena related to vocal pathology, while disregarding spurious information that might confound the [AVQA](#) systems. This procedure should be used with features of different nature, and includes revisiting the complexity characteristics, which are potentially -and theoretically- very informative, but as shown in this thesis, not very consistent.
- Other types of signals or representations related to the speech signal itself should be studied. This includes [ECG](#) obtained through inverse filtering methods, variograms, stroboscopic images, voice and speech range profiles, etc. These signals should be analysed individually and in multimodal scenarios that might contribute to enhance overall performance of [AVQA](#) systems.
- New cross-database experiments should be considered to test out the performance of the automatic detection and assessment systems with larger and more diverse evaluation datasets.
- Classification systems based on [ANN](#) (Boltzmann machines, deep neural networks, etc.) or in variational Bayes should be studied. These approaches have shown to excel in performance in other related fields, for which a translation to the [AVQA](#) field seems reasonable. In addition to that, new testing using [UBM](#)-based classifiers but in much larger databases containing both normophonic and dysphonic registers is required. This include some testing on ancillary datasets matching the lexical content of the target corpus.
- Other types of ordinal classifiers and regressors should be studied, such as those based on [SVM](#), other types of [ANN](#) or Gaussian processes for ordinal regression.
- It has been shown that [MS](#) constitutes an interesting source of information. For this reason we encourage an in-depth study of the [MS](#) representations of normophonic and dysphonic conditions. The aim is to propose new characteristics that provide hindsight about vocal quality using the information contained in the modulation and frequency bands of the [MS](#).

- An approach consisting in the decomposition of voice and speech signals into components relevant for dysphonic and normophonic speech analysis might be of interest; in such a manner that certain properties describing pathology such as modulation noise, additive noise, or tremor can be studied separately.
- The methodologies proposed in this thesis should be extended to the study of other types of pathologies such as those affecting speech due to neurological disorders (Parkinson's disease, spasmodic dysphonia, etc.). In particular the focus should be put into studying running speech tasks to characterise phenomena related to dysarthria, or to extract relevant information for the characterisation of phonatory and prosodic aspects of disordered speech.
- The study of extralinguistic information should also be extended to the assessment task, as it is reasonable to consider that the evaluator might be biased towards certain types of vocal quality, or to certain characteristics of the speaker itself such as its sex or age.

Part VI

APPENDICES

APPENDIX I: CORRELATION ANALYSIS FOR DIFFERENT
TYPES OF SPEECH MATERIAL.

Table A.1. *Trial D2_HUPA-A*: correlation analysis between all the tested characteristics and their labels, using the [HUPA](#) database and the vowel /a/.

				U_e		NU_e	
		ρ	$\rho_{\mathcal{R}}$			ρ	$\rho_{\mathcal{R}}$
Pert	NNE	0.00	0.28	DynInv	D_2	-0.12	-0.12
	CHNR	0.42	0.40		LLE	-0.29	-0.29
	GNE	0.48	0.49		RPDE	-0.03	-0.03
SpecCeps	PLP	-	0.24	LongRange	h^e	-0.07	-0.07
	MFCC	-	0.20		DFA	-0.10	-0.10
	CPPS	0.25	0.25	Reg	ApEn	0.22	0.22
	LHr	0.07	0.10		SampEn	-0.06	-0.06
MSs	MSH	-0.25	0.24		mSampEn	0.05	0.05
	CIL	0.32	0.31		GSampEn	0.01	0.01
	RALA	0.48	0.48		FuzzyEn	-0.04	-0.04
	MSP ₂₅	0.34	0.34		rHMMEn	-0.17	-0.17
	MSP ₇₅	0.30	0.30	EntAtt	sHMMEn	-0.16	-0.16
	MSP ₉₅	0.33	0.32		PE	-0.18	-0.18
						0.00	0.09

Table A.2. *Trial D2_SVD-A*: correlation analysis between all the tested characteristics and their labels using the [SVD](#) database and the vowel /a/.

				U_e		NU_e	
				ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$
Pert	NNE	-0.01	0.27	DynInv	D_2	0.00	0.20
	CHNR	0.37	0.35		LLE	-0.22	0.23
	GNE	0.34	0.34		RPDE	-0.26	0.26
SpecCeps	PLP	-	0.25	LongRange	h^e	-0.27	0.26
	MFCC	-	0.25		DFA	-0.14	0.19
	CPPS	0.18	0.17	Reg	ApEn	0.07	0.08
	LHr	0.10	0.11		SampEn	-0.02	0.09
MSs	MSH	0.18	0.17		mSampEn	0.01	0.06
	CIL	-0.25	0.24		GSampEn	0.00	0.09
	RALA	-0.32	0.31		FuzzyEn	0.04	0.10
	MSP ₂₅	-0.22	0.21	EntAtt	rHMMEn	-0.10	0.12
	MSP ₇₅	-0.23	0.23		sHMMEn	-0.10	0.11
	MSP ₉₅	-0.25	0.25		PE	-0.24	0.23

Table A.3. *Trial D2_GMar-A*: correlation analysis between all the tested characteristics and their labels using the [GMar](#) database and the vowel /a/.

				U_e		NU_e	
				ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$
Pert	NNE	0.01	0.17	DynInv	D_2	-0.08	0.19
	CHNR	0.33	0.31		LLE	-0.22	0.23
	GNE	0.32	0.32		RPDE	-0.06	0.11
SpecCeps	PLP	-	0.23	LongRange	h^e	-0.13	0.19
	MFCC	-	0.21		DFA	-0.12	0.14
	CPPS	0.23	0.23	Reg	ApEn	-0.08	0.14
	LHr	0.20	0.24		SampEn	-0.08	0.08
MSs	MSH	0.12	0.12		mSampEn	0.02	0.08
	CIL	-0.25	0.24		GSampEn	-0.18	0.18
	RALA	-0.37	0.36		FuzzyEn	-0.18	0.18
	MSP ₂₅	-0.25	0.25	EntAtt	rHMMEn	-0.06	0.08
	MSP ₇₅	-0.24	0.24		sHMMEn	-0.05	0.06
	MSP ₉₅	-0.26	0.25		PE	-0.16	0.17

Table A.4. *Trial D2_SVD-I*: correlation analysis between all the tested characteristics and their labels using the [SVD](#) database and the vowel /i/.

				U_e		NU_e	
				ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$
Pert	NNE	0.00	0.16	DynInv	D_2	-0.01	0.10
	CHNR	0.26	0.26		LLE	0.08	0.10
	GNE	0.26	0.26		h^e	-0.11	0.14
SpecCeps	PLP	-	0.22	LongRange	DFA	-0.14	0.16
	MFCC	-	0.22		RPDE	0.00	0.06
	CPPS	0.12	0.13	Reg	ApEn	-0.12	0.15
	LHr	-0.07	0.07		SampEn	-0.03	0.06
MSs	MSH	0.19	0.20		mSampEn	-0.04	0.09
	CIL	-0.18	0.18		GSampEn	0.16	0.17
	RALA	-0.25	0.26		FuzzyEn	0.00	0.05
	MSP ₂₅	-0.17	0.17		rHMMEn	0.05	0.05
	MSP ₇₅	-0.17	0.16	EntAtt	sHMMEn	0.06	0.07
	MSP ₉₅	-0.16	0.15		PE	-0.15	0.13

Table A.5. *Trial D2_GMar-I*: correlation analysis between all the tested characteristics and their labels using the [GMar](#) database and the vowel /i/.

				U_e		NU_e	
				ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$
Pert	NNE	0.03	0.12	DynInv	D_2	0.00	0.20
	CHNR	0.29	0.28		LLE	-0.22	0.23
	GNE	0.28	0.28		RPDE	-0.26	0.26
SpecCeps	PLP	-	0.25	LongRange	h^e	-0.27	0.26
	MFCC	-	0.20		DFA	-0.14	0.19
	CPP	0.19	0.18	Reg	ApEn	0.07	0.08
	LHr	0.00	0.04		SampEn	-0.02	0.09
MSs	MSH	0.12	0.12		mSampEn	0.01	0.06
	CIL	-0.25	0.24		GSampEn	0.00	0.09
	RALA	-0.37	0.37		FuzzyEn	0.04	0.10
	MSP ₂₅	-0.25	0.25	EntAtt	rHMMEn	-0.10	0.12
	MSP ₇₅	-0.24	0.24		sHMMEn	-0.10	0.11
	MSP ₉₅	-0.26	0.26		PE	-0.24	0.23

Table A.6. *Trial D2_SVD-U*: correlation analysis between all the tested characteristics and their labels using the [SVD](#) database and the vowel /u/.

				U_e				NU_e			
				ρ				$\rho_{\mathcal{R}}$			
				ρ				$\rho_{\mathcal{R}}$			
				ρ				$\rho_{\mathcal{R}}$			
Pert	NNE	0.00	0.09	DynInv	D_2	-0.01	0.13	-0.05	0.04		
	CHNR	0.15	0.14		LLE	-0.10	0.10	0.04	0.05		
	GNE	0.14	0.13		h^e	-0.12	0.13	-0.12	0.13		
SpecCeps	PLP	-	0.19	LongRange	DFA	-0.17	0.17	-	-		
	MFCC	-	0.19		RPDE	0.02	0.08	-	-		
	CPPS	0.13	0.13		ApEn	-0.13	0.13	0.05	0.14		
	LHr	0.03	0.04		SampEn	-0.03	0.04	0.03	0.06		
MSs	MSH	0.09	0.08	Reg	mSampEn	-0.05	0.06	0.06	0.12		
	CIL	-0.04	0.05		GSampEn	0.12	0.13	0.00	0.06		
	RALA	-0.14	0.14		FuzzyEn	0.01	0.05	0.02	0.08		
	MSP ₂₅	-0.03	0.06	EntAtt	rHMMEn	-0.04	0.07	0.01	0.01		
	MSP ₇₅	-0.08	0.07		sHMMEn	-0.05	0.06	0.01	0.01		
	MSP ₉₅	-0.14	0.13		PE	-0.16	0.14	-0.09	0.08		

Table A.7. *Trial D2_GMar-U*: correlation analysis between all the tested characteristics and their labels using the [GMar](#) database and the vowel /u/.

				U_e		NU_e			
				ρ	$\rho_{\mathcal{R}}$	ρ	$\rho_{\mathcal{R}}$		
		ρ	$\rho_{\mathcal{R}}$						
Pert	NNE	0.02	0.13	DynInv	D_2	-0.03	0.10	-0.07	0.08
	CHNR	0.13	0.12		LLE	-0.08	0.09	0.08	0.09
	GNE	0.18	0.19		h^e	-0.02	0.08	-0.02	0.08
SpecCeps	PLP	-	0.22	LongRange	DFA	-0.02	0.05	-	-
	MFCC	-	0.15		RPDE	-0.11	0.12	-	-
	CPPS	0.20	0.21	Reg	ApEn	-0.06	0.06	0.08	0.10
	LHr	0.04	0.08		SampEn	-0.02	0.07	0.07	0.09
MSs	MSH	0.01	0.02		mSampEn	0.18	0.17	0.02	0.05
	CIL	-0.14	0.19		GSampEn	0.06	0.07	-0.08	0.08
	RALA	0.01	0.04		FuzzyEn	-0.04	0.05	-0.07	0.08
	MSP ₂₅	-0.15	0.16	EntAtt	rHMMEn	0.03	0.04	0.03	0.03
	MSP ₇₅	-0.13	0.14		sHMMEn	0.01	0.03	0.02	0.02
MSP ₉₅	-0.14	0.15	PE		-0.13	0.12	-0.14	0.14	

B

APPENDIX II: FEATURE RANKING OF THE ASSESSMENT EXPERIMENTS.

Table B.1. *Experiment A1*: Best features ranked according to the feature selection algorithms [MIM](#), [mRMR](#), [JMI](#) for the partition of the vowel /a/ in [HUPA](#) dataset, and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits.

	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
	MIM	mRMR	JMI	MIM	mRMR	JMI	MIM	mRMR	JMI
Pert	GNE	GNE	GNE	NNE	NNE	NNE	NNE	NNE	NNE
	NNE	CHNR	NNE	GNE	GNE	GNE	GNE	GNE	GNE
	CHNR	NNE	CHNR	CHNR	CHNR	CHNR	CHNR	CHNR	CHNR
SpecCeps	CPPS	CPPS	CPPS	CPPS	CPPS	CPPS	CPPS	CPPS	CPPS
	PLP(2)	PLP(3)	PLP(3)	PLP(2)	PLP(3)	MFCC(12)	PLP(3)	PLP(3)	PLP(3)
	PLP(3)	PLP(9)	PLP(2)	MFCC(11)	PLP(2)	PLP(6)	PLP(2)	PLP(2)	LHr
	MFCC(2)	LHr	MFCC(1)	PLP(3)	MFCC(1)	PLP(7)	MFCC(2)	MFCC(1)	MFCC(5)
	PLP(9)	MFCC(10)	MFCC(2)	MFCC(12)	PLP(11)	LHr	MFCC(11)	MFCC(2)	MFCC(8)
MS	RALA	RALA	RALA	RALA	RALA	RALA	RALA	RALA	RALA
	MSP ₂₅	MSH	CIL	MSP ₂₅	CIL	MSH	MSP ₂₅	MSP ₉₅	MSH
	CIL	MSP ₇₅	MSP ₉₅	MSP ₉₅	MSP ₉₅	MSP ₇₅	MSP ₇₅	CIL	MSP ₇₅
	MSP ₉₅	MSP ₉₅	MSH	CIL	MSH	MSP ₉₅	MSP ₉₅	MSH	MSP ₉₅
	MSP ₇₅	MSP ₂₅	MSP ₂₅	MSP ₇₅	MSP ₇₅	MSP ₂₅	CIL	MSP ₂₅	MSP ₂₅
	MSH	CIL	MSP ₇₅	MSH	MSP ₂₅	CIL	MSH	MSP ₇₅	CIL
U_e	PE	PE	PE	PE	PE	PE	PE	PE	PE
	LLE	SampEn	DFA	LLE	DFA	D_2	LLE	DFA	D_2
	RPDE	D_2	mSampEn	ApEn	LLE	LLE	RPDE	LLE	LLE
	ApEn	LLE	h^e	mSampEn	h^e	ApEn	ApEn	h^e	SampEn
	mSampEn	sHMMEn	LLE	DFA	mSampEn	sHMMEn	DFA	mSampEn	sHMMEn
NU_e	PE	PE	PE	LLE	LLE	LLE	PE	PE	PE
	LLE	h^e	DFA	PE	DFA	h^e	LLE	DFA	h^e
	D_2	rHMMEn	LLE	DFA	ApEn	rHMMEn	D_2	LLE	rHMMEn
	ApEn	RPDE	ApEn	h^e	PE	ApEn	DFA	ApEn	RPDE
	SampEn	ApEn	FuzzyEn	ApEn	h^e	RPDE	ApEn	FuzzyEn	ApEn

Table B.2. *Experiment A1*: Best features ranked according to the feature selection algorithms [MIM](#), [mRMR](#), [JMI](#) for the partition of the vowel /a/ in [SVD](#) dataset, and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits.

	$\mathcal{G}$			$\mathcal{B}$			$\mathcal{R}$		
	MIM	mRMR	JMI	MIM	mRMR	JMI	MIM	mRMR	JMI
Pert	NNE	NNE	NNE	NNE	NNE	NNE	CHNR	CHNR	CHNR
	CHNR	GNE	GNE	GNE	GNE	GNE	NNE	GNE	GNE
	GNE	CHNR	CHNR	CHNR	CHNR	CHNR	GNE	NNE	NNE
SpecCeps	PLP(3)	PLP(3)	PLP(3)	PLP(3)	PLP(3)	PLP(3)	LHr	LHr	LHr
	MFCC(2)	CPPS	PLP(2)	PLP(2)	PLP(2)	LHr	PLP(3)	PLP(3)	MFCC(2)
	LHr	LHr	MFCC(1)	LHr	MFCC(1)	CPPS	MFCC(2)	MFCC(16)	CPPS
	PLP(2)	PLP(5)	MFCC(2)	MFCC(2)	MFCC(2)	PLP(5)	PLP(2)	PLP(2)	PLP(5)
	CPPS	MFCC(16)	CPPS	CPPS	CPPS	MFCC(7)	MFCC(16)	MFCC(2)	MFCC(8)
MS	RALA	RALA	RALA	RALA	RALA	RALA	MSP ₉₅	MSP ₉₅	MSP ₉₅
	CIL	MSH	CIL	MSP ₉₅	MSP ₉₅	MSH	CIL	RALA	MSH
	MSP ₉₅	MSP ₇₅	MSP ₉₅	MSP ₂₅	CIL	MSP ₇₅	MSP ₇₅	CIL	RALA
	MSP ₇₅	MSP ₂₅	MSH	CIL	MSH	MSP ₂₅	RALA	MSH	MSP ₂₅
	MSP ₂₅	MSP ₉₅	MSP ₇₅	MSP ₇₅	MSP ₇₅	MSP ₉₅	MSP ₂₅	MSP ₇₅	MSP ₇₅
	MSH	CIL	MSP ₂₅	MSH	MSP ₂₅	CIL	MSH	MSP ₂₅	CIL
$\mathcal{U}_e$	PE	PE	PE	PE	PE	PE	PE	PE	PE
	LLE	h^e	h^e	LLE	h^e	h^e	RPDE	h^e	D_2
	RPDE	LLE	RPDE	RPDE	LLE	LLE	LLE	RPDE	LLE
	h^e	D_2	DFA	h^e	DFA	ApEn	DFA	LLE	SampEn
	DFA	SampEn	LLE	DFA	RPDE	D_2	h^e	DFA	RPDE
$\mathcal{N}\mathcal{U}_e$	PE	PE	PE	PE	PE	PE	PE	PE	PE
	h^e	h^e	LLE	ApEn	LLE	h^e	LLE	LLE	h^e
	ApEn	rHMMEn	DFA	h^e	DFA	rHMMEn	RPDE	DFA	rHMMEn
	DFA	ApEn	ApEn	FuzzyEn	ApEn	ApEn	DFA	ApEn	RPDE
	LLE	RPDE	FuzzyEn	LLE	FuzzyEn	RPDE	h^e	RPDE	ApEn

Table B.3. *Experiment A1*: Best features ranked according to the feature selection algorithms [MIM](#), [mRMR](#), [JMI](#) for the partition of the vowel /a/ in [GMar](#) dataset, and the $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{B}$ traits.

	$\mathcal{G}$			$\mathcal{R}$			$\mathcal{B}$		
	MIM	mRMR	JMI	MIM	mRMR	JMI	MIM	mRMR	JMI
Pert	NNE	NNE	NNE	GNE	GNE	GNE	GNE	GNE	GNE
	GNE	GNE	GNE	NNE	NNE	CHNR	NNE	NNE	CHNR
	CHNR	CHNR	CHNR	CHNR	CHNR	NNE	CHNR	CHNR	NNE
SpecCeps	PLP(2)	PLP(2)	PLP(2)	PLP(2)	PLP(2)	PLP(2)	PLP(3)	PLP(3)	PLP(3)
	LHr	PLP(3)	PLP(3)	LHr	PLP(3)	PLP(9)	PLP(2)	PLP(2)	CPPS
	PLP(3)	MFCC(19)	MFCC(19)	MFCC(1)	MFCC(1)	PLP(7)	LHr	MFCC(16)	LHr
	CPPS	CPPS	PLP(8)	PLP(7)	MFCC(5)	PLP(3)	MFCC(19)	MFCC(19)	MFCC(19)
	MFCC(1)	MFCC(5)	MFCC(1)	PLP(3)	PLP(9)	CPPS	MFCC(16)	MFCC(2)	MFCC(5)
MS	RALA	RALA	RALA	RALA	RALA	RALA	RALA	RALA	RALA
	CIL	MSP ₇₅	CIL	MSP ₉₅	CIL	MSP ₇₅	MSP ₉₅	CIL	MSH
	MSP ₉₅	MSH	MSP ₇₅	CIL	MSP ₉₅	MSH	CIL	MSP ₉₅	MSP ₉₅
	MSP ₇₅	MSP ₉₅	MSP ₉₅	MSP ₂₅	MSP ₇₅	MSP ₉₅	MSP ₂₅	MSP ₇₅	MSP ₂₅
	MSP ₂₅	MSP ₂₅	MSP ₂₅	MSP ₇₅	MSP ₂₅	MSP ₂₅	MSP ₇₅	MSP ₂₅	MSP ₇₅
	MSH	CIL	MSH	MSH	MSH	CIL	MSH	MSH	CIL
U_e	PE	PE	PE	PE	PE	PE	PE	PE	PE
	RPDE	D_2	DFA	LLE	DFA	D_2	RPDE	DFA	D_2
	DFA	SampEn	h^e	h^e	mSampEn	LLE	DFA	h^e	SampEn
	h^e	LLE	RPDE	ApEn	h^e	SampEn	h^e	RPDE	LLE
	LLE	mSampEn	mSampEn	DFA	LLE	ApEn	LLE	mSampEn	sHMMEn
NU_e	LLE	LLE	LLE	LLE	LLE	LLE	LLE	LLE	LLE
	ApEn	h^e	ApEn	h^e	DFA	h^e	DFA	DFA	h^e
	DFA	rHMMEn	DFA	ApEn	FuzzyEn	rHMMEn	h^e	FuzzyEn	rHMMEn
	h^e	ApEn	FuzzyEn	PE	h^e	ApEn	ApEn	h^e	mSampEn
	FuzzyEn	RPDE	h^e	FuzzyEn	ApEn	RPDE	FuzzyEn	ApEn	SampEn

REFERENCES

- [1] J. A. Seikel, D. G. Drumright; and D. W. King, *Anatomy & Physiology for Speech, Language, and Hearing*. Cengage Learning, 2010, p. 832.
- [2] J. Clark, C. Yallop, and J. Fletcher, *An Introduction to Phonetics and Phonology*. Blackwell Publishing, 2007, p. 504.
- [3] T. J. Hixon, G. Weismer, and J. D. Hoit, *Preclinical Speech Science: Anatomy, Physiology, Acoustics, and Perception*, Second. Plural Publishing, 2013, p. 712.
- [4] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, United States ed. PTR Prentice Hall, 1993.
- [5] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education, 2008.
- [6] D. M. Howard and J. A. Angus, *Acoustics and Psychoacoustics*. Focal Press, 2009.
- [7] Tavin, *Sketch of the human vocal tract*, [Online; accessed 15-July-2016], 2011. [Online]. Available: <https://commons.wikimedia.org/wiki/File:VocalTract.svg>.
- [8] A. Behrman, *Speech and voice science*. Plural Publishing Inc, 2007.
- [9] Wugapodes, *A simple, unlabeled diagram of the human vocal tract*. [Online; accessed 16-August-2016], 2015. [Online]. Available: https://commons.wikimedia.org/wiki/File:Midsagittal_diagram_unlabeled.svg.
- [10] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. (Huang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.
- [11] L. R. Rabiner and R. W. Schafer, «Introduction to digital speech processing,» *Foundations and trends in signal processing*, vol. 1, no. 1, pp. 1–194, 2007.
- [12] J. Laver, *Principles of Phonetics*. Cambridge: Cambridge University Press, 1994, p. 736.
- [13] H. Traunmüller, «Conventional, Biological and Environmental Factors in Speech Communication: A Modulation Theory,» *Phonetica*, vol. 51, no. 1-3, pp. 170–183, 1994.
- [14] H. Traunmüller, «Evidence for demodulation in speech perception,» 2000, pp. 790–793.
- [15] A. S.-L.-H. Association, «Definitions of Communication Disorders and Variations,» Tech. Rep., 1993.
- [16] J. B. Snow and J. J. Ballenger, *Ballenger's Otorhinolaryngology Head and Neck Surgery*, B. Decker, Ed. 2003, p. 1616.
- [17] 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.
- [18] M. H. Bahari, M. McLaren, H. Van hamme, and D. a. Van Leeuwen, «Speaker age estimation using i-vectors,» *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99–108, 2014.

- [19] D. Childers and K. Wu, «Gender recognition from speech. Part II: Fine analysis.» *The Journal of the Acoustical Society of America*, vol. 90, no. 4 Pt 1, pp. 1841–56, 1991.
- [20] M. El Ayadi, M. S. Kamel, and F. Karray, «Survey on speech emotion recognition: Features, classification schemes, and databases,» *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
- [21] B. Schuller and G. Rigoll, «Recognising Interest in Conversational Speech – Comparing Bag of Frames and Supra-segmental Features,» in *INTERSPEECH*, 2009, pp. 1999–2002.
- [22] F. Biadisy, «Automatic Dialect and Accent Recognition and its Application to Speech Recognition,» Columbia University, 2011, pp. 1–171.
- [23] J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, S. Aguilera-Navarro, and P. Gómez-Vilda, «An integrated tool for the diagnosis of voice disorders,» *Medical Engineering & Physics*, vol. 28, no. 3, pp. 276–289, 2006.
- [24] C. Moers, B. Möbius, F. Rosanowski, *et al.*, «Vowel- and text-based cepstral analysis of chronic hoarseness,» *Journal of Voice*, vol. 26, no. 4, pp. 416–424, 2012.
- [25] A. E. Aronson and D. Bless, *Clinical voice disorders*, 4th ed. Thieme, 2009.
- [26] N. R. Williams, «Occupational groups at risk of voice disorders: A review of the literature,» *Occupational medicine*, vol. 53, no. 7, pp. 456–460, 2003.
- [27] N. B. Anderson and G. H. Shames, *Human communication disorders: An introduction*, 8th ed. Pearson, Jul. 2010.
- [28] D. R. Boone, S. C. McFarlane, and S. L. V. Berg, *The Voice and Voice Therapy*, 7th ed. Allyn & Bacon, 2004, p. 360.
- [29] R. García-Tapia and I. Cobeta-Marco, *Diagnóstico y tratamiento de la Voz*, Garsi, Ed. 1996, p. 372.
- [30] Prejun, *Vocal cord positions regarding paralyses*, [Online; accessed 3-March-2017], 2011. [Online]. Available: <https://en.wikipedia.org/wiki/File:Vocalcordparalyses.png>.
- [31] C. for Disease Control and Prevention, *Illustration of a baby with a cleft lip*, [Online; accessed 31-January-2018], 2007. [Online]. Available: <https://commons.wikimedia.org/wiki/File:Cleftlip.jpg>.
- [32] C. R. Watts and S. N. Awan, «Use of Spectral/Cepstral Analyses for Differentiating Normal From Hypofunctional Voices in Sustained Vowel and Continuous Speech Contexts,» *Journal of Speech, Language, and Hearing Research*, vol. 54, no. December, pp. 1525–1538, 2011.
- [33] M. Anniko, M. Bernal-Sprekelsen, V. Bonkowsky, P. Bradley, and S. Iurato, *Otorhinolaryngology, Head and Neck Surgery*. Springer, 2010, p. 738.
- [34] S. N. Awan, N. Roy, and S. M. Cohen, «Exploring the relationship between spectral and cepstral measures of voice and the voice handicap index (VHI),» *Journal of Voice*, vol. 28, no. 4, pp. 430–439, 2014.
- [35] J. K. Casper and R. Leonard, *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment*, illustrate. Lippincott Williams & Wilkins, 2006, p. 498.
- [36] S. N. Awan, N. Roy, and C. Dromey, «Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model,» *Clinical linguistics & phonetics*, vol. 23, no. 11, pp. 825–841, 2009.

- [37] 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.
- [38] T. L. Eadie and P. C. Doyle, «Classification of Dysphonic Voice: Acoustic and Auditory-Perceptual Measures,» *Journal of Voice*, vol. 19, no. 1, pp. 1–14, Mar. 2005.
- [39] V. Parsa and D. G. Jamieson, «Acoustic discrimination of pathological voice: sustained vowels versus continuous speech,» *Journal of speech, language, and hearing research : JSLHR*, vol. 44, no. 2, p. 327, 2001.
- [40] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, «Formants of children, women, and men: the effects of vocal intensity variation,» *The Journal of the Acoustical Society of America*, vol. 106, no. 3 Pt 1, pp. 1532–42, 1999.
- [41] L. Baghai-Ravary and S. W. Beet, «Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders,» *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*, no. Fuchs 2005, pp. 7–12, 2013.
- [42] A. Alpan, Y. Maryn, A. Kacha, F. Grenez, and J. Schoentgen, «Multi-band dysperiodicity analyses of disordered connected speech,» *Speech Communication*, vol. 53, no. 1, pp. 131–141, 2011.
- [43] M. Putzer and W. J. Barry, «Instrumental dimensioning of normal and pathological phonation using acoustic measurements,» *Clinical linguistics & phonetics*, vol. 22, no. 6, pp. 407–20, 2008.
- [44] R. D. Kent, *The MIT Encyclopedia of Communication Disorders*. MIT Press, 2004, p. 648.
- [45] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and J. M. Gutiérrez-Arriola, «Characterization of Dysphonic Voices by Means of a Filterbank-Based Spectral Analysis: Sustained Vowels and Running Speech,» *Journal of Voice*, vol. 27, no. 1, pp. 11–23, 2013.
- [46] I. Cobeta, F. Núñez, and S. Fernández, *Patología de la voz*, M. M. Books, Ed. 2013, p. 630.
- [47] P. H. Dejonckere, «Assessment of Voice and Respiratory Function,» in *Surgery of Larynx and Trachea*, Springer Berlin Heidelberg, 2009, pp. 11–26.
- [48] G. B. Kempster, B. R. Gerratt, K. Verdolini 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, p. 124, 2009.
- [49] A. S.-L.-H. Association, [Online; accessed 18-August-2016].
- [50] B. H. Jacobson, A. Johnson, C. Grywalski, *et al.*, «The voice handicap index (VHI),» *American Journal of Speech-Language Pathology*, vol. 6, no. 3, p. 66, Aug. 1997.
- [51] D. Michaelis, M. Fröhlich, and H. W. Strube, «Selection and combination of acoustic features for the description of pathologic voices,» vol. 103, no. 3, pp. 1628–1639, 1998.
- [52] M. Fröhlich, D. Michaelis, H. W. Strube, and E. Kruse, «Acoustic voice analysis by means of the hoarseness diagram,» *Journal of speech, language, and hearing research : JSLHR*, vol. 43, no. 3, pp. 706–20, 2000.

- [53] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguilar, «AHUMADA: A large speech corpus in spanish for speaker characterization and identification,» *Speech Communication*, vol. 31, no. 2-3, pp. 255–264, Jun. 2000.
- [54] S. L. Velleman and B. Z. Pearson, «Differentiating Speech Sound Disorders From Phonological Dialect Differences: Implications for Assessment and Intervention,» *Top Language Disorders*, vol. 30, no. 3, pp. 176–188, 2010.
- [55] C. Huang, T. Chen, S. Z. Li, E. Chang, and J.-L. Zhou, «Analysis of speaker variability,» in *Proc. INTERSPEECH*, 2001, pp. 1377–1380.
- [56] C. Huang, T. Chen, and E. Chang, «Accent Issues in Large Vocabulary Continuous Speech Recognition,» *International Journal of Speech Technology*, vol. 7, no. 2/3, pp. 141–153, 2004.
- [57] M. Benzeghiba, R. De Mori, O. Deroo, *et al.*, «Automatic speech recognition and speech variability: A review,» *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.
- [58] H. Traunmüller and A. Eriksson, «Acoustic effects of variation in vocal effort by men, women, and children,» *The Journal of the Acoustical Society of America*, vol. 107, no. 6, p. 3438, 2000.
- [59] A. L. Rosenthal, S. Y. Lowell, and R. H. Colton, «Aerodynamic and Acoustic Features of Vocal Effort,» *Journal of Voice*, vol. 28, no. 2, pp. 144–153, Mar. 2014.
- [60] S. N. Awan, A. Giovinco, and J. Owens, «Effects of Vocal Intensity and Vowel Type on Cepstral Analysis of Voice,» *Journal of Voice*, vol. 26, no. 5, 670.e15–670.e20, 2012.
- [61] D. Z. Huang, F. D. Minifie, H. Kasuya, and S. X. Lin, «Measures of vocal function during changes in vocal effort level,» *Journal of Voice*, vol. 9, no. 4, pp. 429–438, 1995.
- [62] M. Brockmann, C. Storck, P. N. Carding, and M. J. Drinnan, «Voice Loudness and Gender Effects on Jitter and Shimmer in Healthy Adults,» *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 5, pp. 1152–1160, 2008.
- [63] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, «Reliable Jitter and Shimmer Measurements in Voice Clinics: The Relevance of Vowel, Gender, Vocal Intensity, and Fundamental Frequency Effects in a Typical Clinical Task,» *Journal of Voice*, vol. 25, no. 1, pp. 44–53, 2011.
- [64] P. H. Dejonckere, «Effect of louder voicing on acoustical measurements in dysphonic patients,» *Logopedics Phoniatrics Vocology*, vol. 23, no. 2, pp. 79–84, 1998.
- [65] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, «Speech emotion recognition: Features and classification models,» *Digital Signal Processing*, vol. 22, no. 6, pp. 1154–1160, Dec. 2012.
- [66] M. El Ayadi, M. S. Kamel, and F. Karray, «Survey on speech emotion recognition: Features, classification schemes, and databases,» *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011.
- [67] S. Steidl, A. Batliner, D. Seppi, and B. Schuller, «On the Impact of Children's Emotional Speech on Acoustic and Language Models,» *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 14, pp. 1–14, 2010.
- [68] B. Schuller, J. Stadermann, and G. Rigoll, in *Speech Prosody*, 2006.
- [69] D. Childers and K. Wu, «Automatic recognition of gender by voice,» in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988, pp. 603–606.

- [70] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernández-Cordero, «Gender-dependent phonetic refraction for speaker recognition,» in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, vol. 1, 2002, pp. I–149.
- [71] M. Södersten, S. Hertegård, and B. Hammarberg, «Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women,» *Journal of Voice*, vol. 9, no. 2, pp. 182–197, 1995.
- [72] H. M. Hanson and E. S. Chuang, «Glottal characteristics of male speakers: acoustic correlates and comparison with female data,» *The Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 1064–77, 1999.
- [73] D. H. Klatt and L. C. Klatt, «Analysis, synthesis, and perception of voice quality variations among female and male talkers,» *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.
- [74] K. Wu and D. G. Childers, «Gender recognition from speech. Part I: Coarse analysis,» *The Journal of the Acoustical Society of America*, vol. 90, no. 4 Pt 1, pp. 1828–40, 1991.
- [75] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, «Differences in voice quality between men and women: use of the long-term average spectrum (LTAS),» *Journal of Voice*, vol. 10, no. 1, pp. 59–66, 1996.
- [76] H. M. Hanson, «Glottal characteristics of female speakers: acoustic correlates,» *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–81, 1997.
- [77] E. Holmberg, R. Hillman, and J. Perkell, «Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice,» *The Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1988.
- [78] R. B. Monsen and a. M. Engebretson, «Study of variations in the male and female glottal wave,» *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 981–93, 1977.
- [79] J. Abitbol, P. Abitbol, and B. Abitbol, «Sex hormones and the female voice,» *Journal of Voice*, vol. 13, no. 3, pp. 424–46, 1999.
- [80] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.
- [81] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Wiley-Interscience, 2011, p. 688.
- [82] D. Chan, A. Fourcin, D. Gibbon, et al., *EUROM - A spoken language resource for the EU*, 1995.
- [83] Massachusetts Eye and Ear Infirmary, *Voice disorders database, version.1.03 [cd-rom]*, Lincoln Park, NJ: Kay Elemetrics Corp, 1994.
- [84] V. Parsa and D. G. Jamieson, «Identification of pathological voices using glottal noise measures,» *Journal of speech, language, and hearing research*, vol. 43, no. 2, p. 469, 2000.
- [85] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, «Methodological issues in the development of automatic systems for voice pathology detection,» *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006.

- [86] N. Malyska, T. Quatieri, and D. Sturim, «Automatic Dysphonia Recognition using Biologically-Inspired Amplitude-Modulation Features,» in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, IEEE, 2005, pp. 873–876.
- [87] *Saarbrücken voice database*. [Online]. Available: <http://www.stimmdatenbank.coli.uni-saarland.de/index.php4>.
- [88] A. Al-nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, «Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions,» *Journal of Voice*, vol. 31, no. 1, pp. 3–15, 2017.
- [89] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, «An improved method for voice pathology detection by means of a HMM-based feature space transformation,» *Pattern Recognition*, vol. 43, no. 9, pp. 3100–3112, Sep. 2010.
- [90] C. D. P. Crovato and A. Schuck, «The Use of Wavelet Packet Transform and Artificial Neural Networks in Analysis and Classification of Dysphonic Voices,» *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 10, pp. 1898–1900, 2007.
- [91] R. Behroozmand and F. Almasganj, «Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis,» *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 474–485, Apr. 2007.
- [92] C. Fabris, W. De Colle, and G. Sparacino, «Voice disorders assessed by (cross-) Sample Entropy of electroglottogram and microphone signals,» *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 920–926, Nov. 2013.
- [93] L. A. Forero M., M. Kohler, M. M. Vellasco, and E. Cataldo, «Analysis and Classification of Voice Pathologies Using Glottal Signal Parameters,» *Journal of Voice*, vol. 30, no. 5, pp. 549–556, Sep. 2016.
- [94] A. Giovanni, M. Ouaknine, and J. M. Triglia, «Determination of largest Lyapunov exponents of vocal signal: application to unilateral laryngeal paralysis,» *Journal of voice : Official journal of the Voice Foundation*, vol. 13, no. 3, pp. 341–54, Sep. 1999.
- [95] S. Hadjitodorov, B. Boyanov, and B. Teston, «Laryngeal pathology detection by means of class-specific neural maps,» *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 1, pp. 68–73, 2000.
- [96] J. Jiang and Y. Zhang, «Nonlinear dynamic analysis of speech from pathological subjects,» *Electronics Letters*, vol. 38, no. 6, pp. 294–295, 2002.
- [97] «A novel hybrid of genetic algorithm and ANN for developing a high efficient method for vocal fold pathology diagnosis,» *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 3, Dec. 2015.
- [98] P. Mitev and S. Hadjitodorov, «A method for turbulent noise estimation in voiced signals,» *Med. Biol. Eng. Comput*, vol. 38, pp. 625–631, 2000.
- [99] M. Petrović-Lazić, S. Babac, M. Vuković, R. Kosanović, and Z. Ivanković, «Acoustic Voice Analysis of Patients With Vocal Fold Polyp,» *Journal of Voice*, vol. 25, no. 1, pp. 94–97, Jan. 2011.
- [100] P. Saidi and F. Almasganj, «Voice Disorder Signal Classification Using M-Band Wavelets and Support Vector Machine,» *Circuits, Systems, and Signal Processing*, vol. 34, pp. 2727–2738, 2015.

- [101] P. R. Scalassara, M. E. Dajer, C. D. Maciel, R. C. Guido, and J. C. Pereira, «Relative entropy measures applied to healthy and pathological voice characterization,» *Applied Mathematics and Computation*, vol. 207, no. 1, pp. 95–108, Jan. 2009.
- [102] V. Uloza, A. Verikas, M. Bacauskiene, *et al.*, «Categorizing normal and pathological voices: Automated and perceptual categorization,» *Journal of Voice*, vol. 25, no. 6, pp. 700–708, 2011.
- [103] E. Vaiciukynas, A. Verikas, A. Gelzinis, *et al.*, «Fusion of voice signal information for detection of mild laryngeal pathology,» *Applied Soft Computing*, vol. 18, pp. 91–103, May 2014.
- [104] J. H. Van Stan, D. D. Mehta, S. M. Zeitels, *et al.*, «Average ambulatory measures of sound pressure level, fundamental frequency, and vocal dose do not differ between adult females with phonotraumatic lesions and matched control subjects,» *Annals of Otology, Rhinology and Laryngology*, vol. 124, no. 11, pp. 864–874, 2015.
- [105] A. Verikas, A. Gelzinis, E. Vaiciukynas, *et al.*, «Data dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: Acoustic versus contact microphone,» *Medical Engineering & Physics*, vol. 37, no. 2, pp. 210–218, Feb. 2015.
- [106] M. N. Vieira, F. R. McInnes, and M. a. Jack, «On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures,» *The Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1045–1055, 2002.
- [107] Y. Zhang and J. J. Jiang, «Nonlinear dynamic analysis in signal typing of pathological human voices,» *Electronics Letters*, vol. 39, no. 13, p. 1021, 2003.
- [108] P. Henriquez, J. B. Alonso-Hernandez, M. A. Ferrer, *et al.*, «Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics,» *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1186–1195, 2009.
- [109] M. de Oliveira Rosa, J. Pereira, and M. Grellet, «Adaptive estimation of residue signal for voice pathology diagnosis,» *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 1, pp. 96–104, 2000.
- [110] L. Landini, C. Manfredi, V. Positano, M. Santarelli, and N. Vanello, «Non-linear prediction for oesophageal voice analysis,» *Medical Engineering & Physics*, vol. 24, no. 7-8, pp. 529–533, Sep. 2002.
- [111] C. M. Travieso, J. B. Alonso, J. Orozco-Arroyave, *et al.*, «Detection of different voice diseases based on the nonlinear characterization of speech signals,» *Expert Systems with Applications*, vol. 82, pp. 184–195, Oct. 2017.
- [112] V. Péan, M. Ouayoun, and B. C. Fugain, «A Fractal Approach to Normal and Pathological Voices,» *Acta Oto-Laryngologica*, vol. 120, no. 2, pp. 222–224, Jan. 2000.
- [113] V. Majidnezhad, «A HTK-based Method for Detecting Vocal Fold Pathology,» *Acta Informatica Medica*, vol. 22, no. 4, p. 246, 2014.
- [114] S. N. Awan, N. Roy, D. Zhang, and S. M. Cohen, «Validation of the Cepstral Spectral Index of Dysphonia (CSID) as a Screening Tool for Voice Disorders: Development of Clinical Cutoff Scores,» *Journal of Voice*, vol. 30, no. 2, pp. 130–144, Mar. 2016.

- [115] S. Y. Lowell, R. H. Colton, R. T. Kelley, and Y. C. Hahn, «Spectral- and Cepstral-Based Measures During Continuous Speech: Capacity to Distinguish Dysphonia and Consistency Within a Speaker,» *Journal of Voice*, vol. 25, no. 5, e223–e232, 2010.
- [116] D. D. Mehta, J. H. Van Stan, M. Zanartu, *et al.*, «Using Ambulatory Voice Monitoring to Investigate Common Voice Disorders: Research Update.,» *Frontiers in bioengineering and biotechnology*, vol. 3, no. October, p. 155, 2015.
- [117] E. Ma, J. Robertson, C. Radford, *et al.*, «Reliability of Speaking and Maximum Voice Range Measures in Screening for Dysphonia,» *Journal of Voice*, vol. 21, no. 4, pp. 397–406, Jul. 2007.
- [118] M. A. Little, D. A. Costello, and M. L. Harries, «Objective Dysphonia Quantification in Vocal Fold Paralysis: Comparing Nonlinear With Classical Measures,» *Journal of Voice*, vol. 25, no. 1, pp. 21–31, 2011.
- [119] L. F. Brinca, A. P. F. Batista, A. I. Tavares, I. C. Gonçalves, and M. L. Moreno, «Use of Cepstral Analyses for Differentiating Normal From Dysphonic Voices: A Comparative Study of Connected Speech Versus Sustained Vowel in European Portuguese Female Speakers,» *Journal of Voice*, vol. 28, no. 3, pp. 282–286, May 2014.
- [120] I. Esmaili, N. J. Dabanloo, and M. Vali, «Automatic classification of speech dysfluencies in continuous speech based on similarity measures and morphological image processing tools,» *Biomedical Signal Processing and Control*, vol. 23, pp. 104–114, Jan. 2016.
- [121] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, *et al.*, «Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia),» in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 2005)*, 2005, pp. 149–152.
- [122] F. Klingholtz, «Acoustic recognition of voice disorders: a comparative study of running speech versus sustained vowels,» *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2218–2224, 1990.
- [123] J.-W. Lee, S. Kim, and H.-G. Kang, «Detecting pathological speech using contour modeling of harmonic-to-noise ratio,» in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2014, pp. 5969–5973.
- [124] T. Lee, Y. Liu, P.-W. Huang, *et al.*, «Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech,» in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2016-May, IEEE, Mar. 2016, pp. 6475–6479.
- [125] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, and M. De Bodt, «Toward Improved Ecological Validity in the Acoustic Measurement of Overall Voice Quality: Combining Continuous Speech and Sustained Vowels,» *Journal of Voice*, vol. 24, no. 5, pp. 540–555, Sep. 2010.
- [126] C. Middag, Y. Saeys, and J.-p. Martens, «Towards an ASR-free objective analysis of pathological speech,» in *Interspeech*, 2010, pp. 294–297.
- [127] K. Manickam, C. Moore, T. Willard, and N. Slevin, «Quantifying aberrant phonation using approximate entropy in electrolaryngography,» *Speech Communication*, vol. 47, no. 3, pp. 312–321, Nov. 2005.

- [128] C. Moore, K. Manickam, T. Willard, *et al.*, «Spectral pattern complexity analysis and the quantification of voice normality in healthy and radiotherapy patient groups,» *Medical Engineering & Physics*, vol. 26, no. 4, pp. 291–301, May 2004.
- [129] E. Vaiciukynas, A. Verikas, A. Gelzinis, *et al.*, «Fusing voice and query data for non-invasive detection of laryngeal disorders,» *Expert Systems with Applications*, vol. 42, no. 22, pp. 8445–8453, 2015.
- [130] A. Verikas, A. Gelzinis, M. Bacauskiene, *et al.*, «Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders,» *Artificial Intelligence in Medicine*, vol. 49, no. 1, pp. 43–50, 2010.
- [131] Ö. Eskidere and A. Gürhanlı, «Voice Disorder Classification Based on Multitaper Mel Frequency Cepstral Coefficients Features,» *Computational and Mathematical Methods in Medicine*, vol. 2015, pp. 1–12, 2015.
- [132] A. A. Dibazar, T. W. Berger, and S. S. Narayanan, «Pathological Voice Assessment,» in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Aug. 2006, pp. 1669–1673.
- [133] A. Alpan, J. Schoentgen, Y. Maryn, and F. Grenez, «Automatic perceptual categorization of disordered connected speech,» in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 2574–2577.
- [134] K. Shama, A. Krishna, and N. U. Cholaayya, «Study of Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology,» *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 085 286, 2007.
- [135] C. Manfredi, M. D'Aniello, P. Bruscaaglioni, and A. Ismaelli, «A comparative analysis of fundamental frequency estimation methods with application to pathological voices,» *Medical Engineering & Physics*, vol. 22, no. 2, pp. 135–147, Mar. 2000.
- [136] M. Markaki and Y. Stylianou, «Voice Pathology Detection and Discrimination Based on Modulation Spectral Features,» *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.
- [137] G. Muhammad, M. Alsulaiman, Z. Ali, *et al.*, «Voice pathology detection using interlaced derivative pattern on glottal source excitation,» *Biomedical Signal Processing and Control*, vol. 31, pp. 156–164, Jan. 2017.
- [138] S. Jothilakshmi, «Automatic system to detect the type of voice pathology,» *Applied Soft Computing*, vol. 21, pp. 244–249, Aug. 2014.
- [139] R. Wielgat, T. P. Zieliński, T. Woźniak, S. Grabias, and D. Król, «Automatic Recognition of Pathological Phoneme Production,» *Folia Phoniatrica et Logopaedica*, vol. 60, no. 6, pp. 323–331, Nov. 2008.
- [140] S. Hadjitodorov and P. Mitev, «A computer system for acoustic analysis of pathological voices and laryngeal diseases screening,» *Medical Engineering & Physics*, vol. 24, no. 6, pp. 419–429, 2002.
- [141] M. K. Arjmandi and M. Pooyan, «An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine,» *Biomedical Signal Processing and Control*, vol. 7, no. 1, pp. 3–19, Jan. 2012.

- [142] D. Panek, A. Skalski, and J. Gajda, «Quantification of Linear and Non-linear Acoustic Analysis Applied to Voice Pathology Detection,» in *Information Technologies in Biomedicine*, ser. Advances in Intelligent Systems and Computing, E. Piętko, J. Kawa, and W. Wieclawek, Eds., vol. 284, Cham: Springer International Publishing, 2014.
- [143] Y. Qi, R. E. Hillman, and C. Milstein, «The estimation of signal-to-noise ratio in continuous speech for disordered voices,» *Journal of the Acoustical Society of America*, vol. 105, no. 4, pp. 2532–2535, 1999.
- [144] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, «Automatic Voice Pathology Detection With Running Speech by Using Estimation of Auditory Spectrum and Cepstral Coefficients Based on the All-Pole Model,» *Journal of Voice*, vol. 30, no. 6, 757.e7–757.e19, Nov. 2016.
- [145] D. Panek, A. Skalski, J. Gajda, and R. Tadeusiewicz, «Acoustic Analysis Assessment in Speech Pathology Detection,» *International Journal of Applied Mathematics and Computer Science*, vol. 25, pp. 631–643, 2015.
- [146] M. S. Hossain, «Cloud-Supported Cyber-Physical Localization Framework for Patients Monitoring,» *IEEE Systems Journal*, vol. 11, no. 1, pp. 118–127, Mar. 2017.
- [147] J. I. Godino-Llorente, P. Gómez-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 bio-medical engineering*, vol. 53, no. 10, pp. 1943–1953, 2006.
- [148] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, and V. Uloza, «Exploring similarity-based classification of larynx disorders from human voice,» *Speech Communication*, vol. 54, no. 5, pp. 601–610, 2012.
- [149] R. Moran, R. Reilly, P. de Chazal, and P. Lacy, «Telephony-Based Voice Pathology Assessment Using Automated Speech Analysis,» *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, 2006.
- [150] A. Maier, T. Haderlein, U. Eysholdt, *et al.*, «PEAKS - A system for the automatic evaluation of voice and speech disorders,» *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.
- [151] G. Muhammad, T. a. Mesallam, K. H. Malki, *et al.*, «Multidirectional regression (MDR)-based features for automatic voice disorder detection,» *Journal of Voice*, vol. 26, no. 6, 817.e19–817.e27, 2012.
- [152] X. Wang, J. Zhang, and Y. Yan, «Discrimination Between Pathological and Normal Voices Using GMM-SVM Approach,» *Journal of Voice*, vol. 25, no. 1, pp. 38–43, 2011.
- [153] J. Mekyska, E. Janousova, P. Gómez-Vilda, *et al.*, «Robust and complex approach of pathological speech signal analysis,» *Neurocomputing*, vol. 167, pp. 94–111, 2015.
- [154] M. Hariharan, K. Polat, and S. Yaacob, «A new feature constituting approach to detection of vocal fold pathology,» *International Journal of Systems Science*, vol. 45, no. 8, pp. 1622–1634, 2013.
- [155] C. Manfredi, «Adaptive noise energy estimation in pathological speech signals,» *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 11, pp. 1538–1543, 2000.
- [156] J. B. Alonso-Hernandez, J. De Leon, I. Alonso, and M. A. Ferrer, «Automatic detection of pathologies in the voice by HOS based parameters,» *Eurasip Journal on Applied Signal Processing*, vol. 2001, no. 4, pp. 275–284, 2001.

- [157] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, *et al.*, «The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders,» *Journal of Voice*, vol. 24, no. 1, pp. 47–56, 2010.
- [158] A. Dibazar, S. Narayanan, and T. Berger, «Feature analysis for automatic detection of pathological speech,» in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society* [Engineering in Medicine and Biology, vol. 1, IEEE, 2002, pp. 182–183.
- [159] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, «On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices,» *Logopedics, phoniatrics, vocology*, vol. 36, no. 2, pp. 60–69, 2011.
- [160] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolaños, J. D. Arias-Londoño, *et al.*, «Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases,» *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1820–1828, 2015.
- [161] T. Dubuisson, T. Dutoit, B. Gosselin, and M. Remacle, «On the Use of the Correlation between Acoustic Descriptors for the Normal/Pathological Voices Discrimination,» *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 173 967, 2009.
- [162] J. I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda, «Automatic detection of voice impairments from text-dependent running speech,» *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 176–182, 2009.
- [163] J. W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, «An Investigation of Vocal Tract Characteristics for Acoustic Discrimination of Pathological Voices,» *BioMed Research International*, vol. 2013, pp. 1–11, 2013.
- [164] M. Vasilakis and Y. Stylianou, «Spectral jitter modeling and estimation,» *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 183–193, 2009.
- [165] J. I. Godino-Llorente and P. Gomez-Vilda, «Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors,» *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.
- [166] M. Vasilakis and Y. Stylianou, «Voice pathology detection based on short-term jitter estimations in running speech,» *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 153–170, 2009.
- [167] C. Moore, K. Manickam, and N. Slevin, «Collective spectral pattern complexity analysis of voicing in normal males and larynx cancer patients following radiotherapy,» *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 113–119, Apr. 2006.
- [168] G. Daza-Santacoloma, J. D. Arias-Londono, J. I. Godino-Llorente, *et al.*, «Dynamic Feature Extraction: an Application To Voice Pathology Detection,» *Intelligent Automation and Soft Computing*, vol. 15, no. 4, pp. 667–682, 2009.
- [169] 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.
- [170] H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, «Hierarchical Classification and System Combination for Automatically Identifying Physiological and Neuromuscular Laryngeal Pathologies,» *Journal of Voice*, Oct. 2016.

- [171] T. Drugman, T. Dubuisson, and T. Dutoit, «Phase-based information for voice pathology detection,» in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2011, pp. 4612–4615.
- [172] T. Drugman, T. Dubuisson, and T. Dutoit, «On the mutual information between source and filter contributions for voice pathology detection,» in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 1463–1466.
- [173] P. Gómez-Vilda, R. Fernández-Baillo, A. Nieto, *et al.*, «Evaluation of Voice Pathology Based on the Estimation of Vocal Fold Biomechanical Parameters,» *Journal of Voice*, vol. 21, no. 4, pp. 450–476, Jul. 2007.
- [174] C. Carmona-Duarte, R. Plamondon, P. Gómez-Vilda, *et al.*, «Application of the Lognormal Model to the Vocal Tract Movement to Detect Neurological Diseases in Voice,» in *Smart Innovation, Systems and Technologies*, vol. 60, 2016, pp. 25–35.
- [175] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar-Biarge, *et al.*, «Glottal Source biometrical signature for voice pathology detection,» *Speech Communication*, vol. 51, no. 9, pp. 759–781, Sep. 2009.
- [176] J. Godino-Llorente, S. Aguilera-Navarro, and P. Gomez-Vilda, «Automatic detection of voice impairments due to vocal misuse by means of Gaussian mixture models,» in *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2001, pp. 1723–1726.
- [177] A. Roviroso, E. Martínez-Celdrán, A. Ortega, *et al.*, «Acoustic analysis after radiotherapy in T1 vocal cord carcinoma: a new approach to the analysis of voice quality,» *International Journal of Radiation Oncology*Biophysics*, vol. 47, no. 1, pp. 73–79, Apr. 2000.
- [178] V. Uloza, V. Saferis, and I. Uloziene, «Perceptual and Acoustic Assessment of Voice Pathology and the Efficacy of Endolaryngeal Phonomicrosurgery,» *Journal of Voice*, vol. 19, no. 1, pp. 138–145, Mar. 2005.
- [179] M. Döllinger, M. Kunduk, M. Kaltenbacher, *et al.*, «Analysis of vocal fold function from acoustic data simultaneously recorded with high-speed endoscopy,» *Journal of Voice*, vol. 26, no. 6, pp. 726–733, 2012.
- [180] Z. Ali, M. Alsulaiman, I. Elamvazuthi, *et al.*, «Voice pathology detection based on the modified voice contour and SVM,» *Biologically Inspired Cognitive Architectures*, vol. 15, pp. 10–18, Jan. 2016.
- [181] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. Cheyne, and R. E. Hillman, «Mobile Voice Health Monitoring Using a Wearable Accelerometer Sensor and a Smartphone Platform,» *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 11, pp. 3090–3096, Nov. 2012.
- [182] a. M. Sulter, H. P. Wit, H. K. Schutte, and D. G. Miller, «A structured approach to voice range profile (phonetogram) analysis.,» *Journal of speech and hearing research*, vol. 37, no. 5, pp. 1076–85, 1994.
- [183] A. Behrman, C. J. Agresti, E. Blumstein, and G. Sharma, «Meaningful features of voice range profiles from patients with organic vocal fold pathology: A preliminary study,» *Journal of Voice*, vol. 10, no. 3, pp. 269–283, 1996.
- [184] A. E. Hallin, K. Fröst, E. B. Holmberg, and M. Södersten, «Voice and speech range profiles and Voice Handicap Index for males — methodological issues and data,» *Logopedics Phoniatrics Vocology*, vol. 37, no. 2, pp. 47–61, 2012.

- [185] J. Goddard, G. Schlotthauer, M. Torres, and H. Rufiner, «Dimensionality reduction for visualization of normal and pathological speech data,» *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 194–201, Jul. 2009.
- [186] G. Schlotthauer, M. E. Torres, and M. C. Jackson-Menaldi, «A Pattern Recognition Approach to Spasmodic Dysphonia and Muscle Tension Dysphonia Automatic Classification,» *Journal of Voice*, vol. 24, no. 3, pp. 346–353, May 2010.
- [187] M. Shu, J. J. Jiang, and M. Willey, «The Effect of Moving Window on Acoustic Analysis,» *Journal of Voice*, vol. 30, no. 1, pp. 5–10, Jan. 2016.
- [188] V. Uloza, E. Padervinskis, A. Vegiene, *et al.*, «Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening,» *European Archives of Oto-Rhino-Laryngology*, vol. 272, no. 11, pp. 3391–3399, 2015.
- [189] Y. Zhang, J. J. Jiang, L. Biazzo, and M. Jorgensen, «Perturbation and non-linear dynamic analyses of voices from patients with unilateral laryngeal paralysis,» *Journal of Voice*, vol. 19, no. 4, pp. 519–528, 2005.
- [190] Y. Zhang and J. J. Jiang, «Acoustic Analyses of Sustained and Running Voices From Patients With Laryngeal Pathologies,» *Journal of Voice*, vol. 22, no. 1, pp. 1–9, 2008.
- [191] Y. Zhang, C. McGilligan, L. Zhou, M. Vig, and J. J. Jiang, «Nonlinear dynamic analysis of voices before and after surgical excision of vocal polyps,» *The Journal of the Acoustical Society of America*, vol. 115, no. 5 Pt 1, pp. 2270–7, 2004.
- [192] I. Titze, «Workshop on Acoustic Voice Analysis,» *National Centre for Voice and Speech, America*, pp. 1–36, 1994.
- [193] P. R. Scalassara, C. D. Maciel, R. C. Guido, *et al.*, «Autoregressive decomposition and pole tracking applied to vocal fold nodule signals,» *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1360–1367, Aug. 2007.
- [194] 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.
- [195] A. Akbari and M. K. Arjmandi, «Employing linear prediction residual signal of wavelet sub-bands in automatic detection of laryngeal pathology,» *Biomedical Signal Processing and Control*, vol. 18, pp. 293–302, 2015.
- [196] Z. Ali, G. Muhammad, and M. F. Alhamid, «An Automatic Health Monitoring System for Patients Suffering from Voice Complications in Smart Cities,» *IEEE Access*, vol. 5, 2017.
- [197] M. Alsulaiman, «Voice pathology assessment systems for dysphonic patients: Detection, classification, and speech recognition,» *IETE Journal of Research*, vol. 60, no. 2, pp. 156–167, 2014.
- [198] J. I. Godino-Llorente, S. Aguilera-Navarro, and P. Gómez-Vilda, «PC, LPCC and MFCC Parameterisation Applied to the Detection of Voice Impairments,» in *Sixth International Conference on Spoken Language Processing*, 2000.
- [199] G. Muhammad, G. Altuwaijri, M. Alsulaiman, *et al.*, «Automatic voice pathology detection and classification using vocal tract area irregularity,» *Biocybernetics and Biomedical Engineering*, vol. 36, no. 2, pp. 309–317, 2016.

- [200] Z. Ali, M. Alsulaiman, G. Muhammad, *et al.*, «Intra- and Inter-database Study for Arabic, English, and German Databases: Do Conventional Speech Features Detect Voice Pathology?» *Journal of Voice*, vol. 31, no. 3, 386.e1–386.e8, May 2017.
- [201] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, «Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients,» *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.
- [202] R. Fraile, N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and C. Fredouille, «Automatic Detection of Laryngeal Pathologies in Records of Sustained Vowels by Means of Mel-Frequency Cepstral Coefficient Parameters and Differentiation of Patients by Sex,» *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 146–152, 2009.
- [203] F. Amara, M. Fezari, and H. Bourouba, «An Improved GMM-SVM System based on Distance Metric for Voice Pathology Detection,» *Appl. Math. Inf. Sci.*, vol. 10, no. 3, pp. 1061–1070, 2016.
- [204] J. I. Godino-Llorente, P. Gómez-Vilda, F. Cruz-Roldán, M. Blanco-Velasco, and R. Fraile, «Pathological Likelihood Index as a Measurement of the Degree of Voice Normality and Perceived Hoarseness,» *Journal of Voice*, vol. 24, no. 6, pp. 667–677, 2010.
- [205] A. Akbari and M. K. Arjmandi, «An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features,» *Biomedical Signal Processing and Control*, vol. 10, no. 1, pp. 209–223, 2014.
- [206] H. Khadivi Heris, B. Seyed Aghazadeh, and M. Nikkhah-Bahrami, «Optimal feature selection for the assessment of vocal fold disorders,» *Computers in Biology and Medicine*, vol. 39, no. 10, pp. 860–868, Oct. 2009.
- [207] J. Nayak, P. Bhat, R. Acharya, and U. Aithal, «Classification and analysis of speech abnormalities,» *ITBM-RBM*, vol. 26, no. 5-6, pp. 319–327, Oct. 2005.
- [208] K. Umapathy and S. Krishnan, «Feature analysis of pathological speech signals using local discriminant bases technique,» *Medical & biological engineering & computing*, vol. 43, no. 4, pp. 457–64, 2005.
- [209] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, «Discrimination of pathological voices using a time-frequency approach,» *IEEE transactions on bio-medical engineering*, vol. 52, no. 3, pp. 421–30, 2005.
- [210] B. Ghoraani and S. Krishnan, «A Joint Time-Frequency and Matrix Decomposition Feature Extraction Methodology for Pathological Voice Classification,» *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–12, 2009.
- [211] B. Ghoraani, K. Umapathy, L. Sugavaneswaran, and S. Krishnan, «Pathological Speech Signal Analysis Using Time-Frequency Approaches,» *Critical Reviews in Biomedical Engineering*, vol. 40, no. 1, pp. 63–95, 2012.
- [212] B. Radish Kumar, J. S. Bhat, and N. Prasad, «Cepstral analysis of voice in persons with vocal nodules,» *Journal of Voice*, vol. 24, no. 6, pp. 651–653, 2010.
- [213] H. Ghasemzadeh, M. Tajik Khass, M. Khalil Arjmandi, and M. Pooyan, «Detection of vocal disorders based on phase space parameters and Lyapunov spectrum,» *Biomedical Signal Processing and Control*, vol. 22, pp. 135–145, Sep. 2015.

- [214] G. Vaziri, F. Almasganj, and R. Behroozmand, «Pathological assessment of patients' speech signals using nonlinear dynamical analysis,» *Computers in Biology and Medicine*, vol. 40, no. 1, pp. 54–63, Jan. 2010.
- [215] J. Arias-Londoño, J. Godino-Llorente, G. Castellanos-Domínguez, N. Sáenz-Lechón, and V. Osma-Ruiz, «Complexity analysis of pathological voices by means of hidden markov entropy measurements,» in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Sep. 2009, pp. 2248–2251.
- [216] R. F. Leonarduzzi, G. a. Alzamendi, G. Schlotthauer, and M. E. Torres, «Wavelet leader multifractal analysis of period and amplitude sequences from sustained vowels,» *Speech Communication*, vol. 72, pp. 1–12, 2015.
- [217] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, «Detection of Voice Pathology using Fractal Dimension in a Multiresolution Analysis of Normal and Disordered Speech Signals,» *Journal of medical systems*, vol. 40, no. 1, p. 20, 2016.
- [218] L. Matassini, R. Hegger, H. Kantz, and C. Manfredi, «Analysis of vocal disorders in a feature space,» *Medical Engineering & Physics*, vol. 22, no. 6, pp. 413–418, Jul. 2000.
- [219] L. Moro-Velázquez, J. A. Gómez-García, 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.
- [220] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, «Voice Pathology Detection Using Modulation Spectrum-Optimized Metrics,» *Frontiers in Bioengineering and Biotechnology*, vol. 4, no. 1, 2016.
- [221] Z. Ali, M. Talha, and M. Alsulaiman, «A Practical Approach: Design and Implementation of a Healthcare Software for Screening of Dysphonic Patients,» *IEEE Access*, vol. 3536, no. c, pp. 1–1, 2017.
- [222] G. Muhammad, M. F. Alhamid, M. S. Hossain, A. S. Almogren, and A. V. Vasilakos, «Enhanced Living by Assessing Voice Pathology Using a Co-Occurrence Matrix,» *Sensors*, vol. 17, no. 2, p. 267, 2017.
- [223] Y. Maryn, M. De Bodt, and N. Roy, «The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders,» *Journal of Communication Disorders*, vol. 43, no. 3, pp. 161–174, May 2010.
- [224] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, «Pathological Voice Analysis and Classification Based on Empirical Mode Decomposition,» in *Development of Multimodal Interfaces: Active Listening and Synchrony. Lecture Notes in Computer Science*, 2010, pp. 364–381.
- [225] M. Kaleem, B. Ghoraani, A. Guergachi, and S. Krishnan, «Pathological speech signal analysis and classification using empirical mode decomposition.,» *Medical & biological engineering & computing*, vol. 51, pp. 811–21, 2013.
- [226] M. Fröhlich, D. Michaelis, H. W. Strube, and E. Kruse, «Acoustic voice quality description : Case studies for different regions of the hoarseness diagram,» *Advances in Quantitative Laryngoscopy, 2nd 'Round Table*, pp. 143–150, 1997.
- [227] A. Fontes, P. Souza, A. Neto, A. Martins, and L. Silveira, «Classification System of Pathological Voices Using Correntropy,» *Mathematical Problems in Engineering*, vol. 2014, pp. 1–7, 2014.

- [228] P. Gómez-Vilda, E. S. Segundo, L. M. Mazaira-Fernández, A. Álvarez-Marquina, and V. Rodellar-Biarge, «Using Dysphonic Voice to Characterize Speaker's Biometry,» *Language and Law / Linguagem e Direito*, vol. 1, no. 2, pp. 42–66, 2014.
- [229] D. Hemmerling, A. Skalski, and J. Gajda, «Voice data mining for laryngeal pathology assessment,» *Computers in Biology and Medicine*, vol. 69, pp. 270–276, Feb. 2016.
- [230] G. Muhammad and M. Melhem, «Pathological voice detection and binary classification using MPEG-7 audio features,» *Biomedical Signal Processing and Control*, vol. 11, no. 1, pp. 1–9, 2014.
- [231] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, and A. Moqarehzadeh, «Identification of voice disorders using long-time features and support vector machine with different feature reduction methods,» *Journal of Voice*, vol. 25, no. 6, e275–e289, 2011.
- [232] M. Hariharan, K. Polat, R. Sindhu, and S. Yaacob, «A hybrid expert system approach for telemonitoring of vocal fold pathology,» *Applied Soft Computing*, vol. 13, no. 10, pp. 4148–4161, Oct. 2013.
- [233] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, *et al.*, «An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification,» *Journal of Voice*, vol. 31, no. 1, 113.e9–113.e18, Jan. 2017.
- [234] S. Shilaskar, A. Ghatol, and P. Chatur, «Medical decision support system for extremely imbalanced datasets,» *Information Sciences*, vol. 384, pp. 205–219, Apr. 2017.
- [235] J. A. Gómez-García, L. Moro-Velázquez, J. I. Godino-Llorente, and C. G. Castellanos-Domínguez, «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*, no. 79, pp. 50–62, 2016.
- [236] T. Bhuta, L. Patrick, and J. D. Garnett, «Perceptual evaluation of voice quality and its correlation with acoustic measurements,» *Journal of Voice*, vol. 18, no. 3, pp. 299–304, 2004.
- [237] S. N. Awan, N. Roy, and J. J. Jiang, «Nonlinear Dynamic Analysis of Disordered Voice: The Relationship Between the Correlation Dimension (D2) and Pre-/Post-Treatment Change in Perceived Dysphonia Severity,» *Journal of Voice*, vol. 24, no. 3, pp. 285–293, 2010.
- [238] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. IEEE, 1999.
- [239] K. Paliwal and K. Wojcicki, «Effect of analysis window duration on speech intelligibility,» *IEEE Signal Processing Letters*, vol. 15, pp. 785–788, 2008.
- [240] J. Schoentgen, «Spectral models of additive and modulation noise in speech and phonatory excitation signals,» *The Journal of the Acoustical Society of America*, vol. 113, no. 1, p. 553, 2003.
- [241] J. Hanquinet, F. Grenez, and J. Schoentgen, «Synthesis of Disordered Voices,» in *International Conference on Non-Linear Speech Processing, NOLISP 2005*, 2006, pp. 231–241.
- [242] M. Farrús and J. Hernando, «Using Jitter and Shimmer in speaker verification,» *IET Signal Processing*, vol. 3, no. November 2008, p. 247, 2009.

- [243] S. Bielałowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke, «Comparison of voice analysis systems for perturbation measurement,» *Journal of Speech & Hearing Research*, vol. 39, no. 1, pp. 126–134, 1996.
- [244] E. Yumoto, W. J. Gould, and T. Baer, «Harmonics to Noise Ratio as hoarseness index of degree of hoarseness,» *Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1549, 1982.
- [245] 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.
- [246] H. Kasuya, «Normalized noise energy as an acoustic measure to evaluate pathologic voice,» *The Journal of the Acoustical Society of America*, vol. 80, no. 5, p. 1329, 1986.
- [247] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, «Acoustic Correlates of Breathy Vocal Quality,» *Journal of Speech Language and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.
- [248] J. Hillenbrand and R. A. Houde, «Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech,» *Journal of Speech Language and Hearing Research*, vol. 39, no. 2, p. 311, 1996.
- [249] S. S. Stevens and J. Volkmann, «The relation of pitch to frequency: A revised scale,» *The American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.
- [250] G. Fant, «Analysis and synthesis of speech processes,» *Manual of phonetics*, vol. 2, pp. 173–277, 1968.
- [251] H. Florian, G. Stemmer, C. Hacker, and F. Brugnara, «Revising Perceptual Linear Prediction (PLP),» *Interspeech 2005*, pp. 2997–3000, 2005.
- [252] H. Hermansky, «Perceptual linear predictive (PLP) analysis of speech,» *The Journal of the Acoustical Society of America*, vol. 87, p. 1738, 1990.
- [253] R. A. Peters and R. N. Strickland, «Image complexity metrics for automatic target recognizers,» in *Automatic Target Recognizer System and Technology Conference*, 1990, pp. 1–17.
- [254] P. Grassberger, «Randomness, Information, and Complexity,» *Complexity*, no. 1, pp. 1–20, 2012.
- [255] M. Costa, A. Goldberger, and C. Peng, «Multiscale entropy analysis of biological signals,» *Physical Review E*, vol. 71, no. 2, pp. 1–18, 2005.
- [256] M. D. Costa, C.-K. Peng, and A. L. Goldberger, «Multiscale Analysis of Heart Rate Dynamics: Entropy and Time Irreversibility Measures,» *Cardiovascular Engineering*, vol. 8, no. 2, pp. 88–93, Jun. 2008.
- [257] A. L. Goldberger, C.-K. Peng, and L. A. Lipsitz, «What is physiologic complexity and how does it change with aging and disease?» *Neurobiology of Aging*, vol. 23, no. 1, pp. 23–26, 2002.
- [258] I. R. Titze, *The myoelastic aerodynamic theory of phonation*, 1st. National Center for Voice and Speech, 2006.
- [259] I. Steinecke and H. Herzel, «Bifurcations in an asymmetric vocal-fold model,» *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1874–84, 1995.
- [260] R. T. Sataloff, *Laryngology (Sataloff's Comprehensive Textbook of Otolaryngology: Head & Neck Surgery)*. Jaypee Brothers, Medical Publishers Pvt. Limited, 2015.

- [261] F. Takens, «Detecting strange attractors in turbulence,» in *Dynamical Systems and Turbulence*, Springer Berlin Heidelberg, 1981, pp. 366–381.
- [262] K. Judd and A. Mees, «Embedding as a modeling problem,» *Physica D: Nonlinear Phenomena*, vol. 120, no. 3-4, pp. 273–286, 1998.
- [263] S. Garcia and J. Almeida, «Nearest neighbor embedding with different time delays,» *Physical Review E*, vol. 71, no. 3, pp. 1–4, 2005.
- [264] Y. Manabe and B. Chakraborty, «A novel approach for estimation of optimal embedding parameters of nonlinear time series by structural learning of neural network,» *Neurocomputing*, vol. 70, no. 7-9, pp. 1360–1371, 2007.
- [265] L. Pecora, L. Moniz, J. Nichols, and T. Carroll, «A unified approach to attractor reconstruction,» *Complex Dynamics in Physiological Systems: From Heart to Brain*, pp. 3–19, 2009.
- [266] M. Ragulskis and K. Lukoseviciute, «Non-uniform attractor embedding for time series forecasting by fuzzy inference systems,» *Neurocomputing*, vol. 72, no. 10-12, pp. 2618–2626, 2009.
- [267] K. Lukoseviciute and M. Ragulskis, «Evolutionary algorithms for the selection of time lags for time series forecasting by fuzzy inference systems,» *Neurocomputing*, vol. 73, no. 10, pp. 2077–2088, 2010.
- [268] L. Uzal, G. Grinblat, and P. Verdes, «Optimal reconstruction of dynamical systems: A noise amplification approach,» *Physical Review E*, vol. 84, no. 1, p. 016 223, 2011.
- [269] J. A. Gómez-García, J. I. Godino-Llorente, and G. Castellanos-Dominguez, «Non uniform Embedding based on Relevance Analysis with reduced computational complexity: Application to the detection of pathologies from biosignal recordings,» *Neurocomputing*, vol. 132, pp. 148–158, 2014.
- [270] J. Xu, «Adapt the mRMR criterion for unsupervised feature selection,» *Advanced Data Mining and Applications*, pp. 111–121, 2010.
- [271] P. Maragos and A. Potamianos, «Fractal dimensions of speech sounds: Computation and application to automatic speech recognition,» *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1925–1932, Mar. 1999.
- [272] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. Cambridge University Press, 2004.
- [273] A. Osborne and A. Provenzale, «Finite correlation dimension for stochastic systems with power-law spectra,» *Physica D: Nonlinear Phenomena*, vol. 35, no. 3, pp. 357–381, 1989.
- [274] S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz, «A regularity statistic for medical data analysis,» *Journal of Clinical Monitoring*, vol. 7, no. 4, pp. 335–345, 1991.
- [275] A. Guerrero and L. Smith, «Towards coherent estimation of correlation dimension,» *Physics Letters A*, vol. 318, no. 4-5, pp. 373–379, Nov. 2003.
- [276] M. S. Taqqu, V. Teverovsky, and W. Willinger, «Estimators for long-range dependence: An empirical study,» *Fractals*, vol. 03, no. 04, pp. 785–798, Dec. 1995.
- [277] K. Hu, P. C. Ivanov, Z. Chen, P. Carpena, and H. Eugene Stanley, «Effect of trends on detrended fluctuation analysis,» vol. 64, no. 1, p. 011 114, Jun. 2001.

- [278] C. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, «Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series,» *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 5, no. 1, pp. 82–87, Mar. 1995.
- [279] F. Kaffashi, R. Foglyano, C. G. Wilson, and K. A. Loparo, «The effect of time delay on Approximate & Sample Entropy calculations,» *Physica D: Nonlinear Phenomena*, vol. 237, no. 23, pp. 3069–3074, 2008.
- [280] 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.
- [281] 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, H2039–49, 2000.
- [282] L. Xu, K. Wang, and L. Wang, «Gaussian kernel approximate entropy algorithm for analyzing irregularity of time-series,» in *Proceedings of 2005 International Conference on Machine Learning and Cybernetics.*, 2005, pp. 5605–5608.
- [283] W. Chen, Z. Wang, H.-B. 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. 266–72, 2007.
- [284] W. Chen, J. Zhuang, W. Yu, and Z. Wang, «Measuring complexity using FuzzyEn, ApEn, and SampEn,» *Medical engineering & physics*, vol. 31, no. 1, pp. 61–8, 2009.
- [285] 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.
- [286] C. Bandt, «Ordinal time series analysis,» *Ecological Modelling*, vol. 182, no. 3–4, pp. 229–238, 2005.
- [287] 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. 12, pp. 1553–1577, 2012.
- [288] J. D. Arias-Londoño and J. I. Godino-Llorente, «Entropies from Markov Models as Complexity Measures of Embedded Attractors,» *Entropy*, vol. 17, no. 6, pp. 3595–3620, 2015.
- [289] M. Cord and P. Cunningham, Eds., *Machine Learning Techniques for Multimedia*, ser. Cognitive Technologies. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.
- [290] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, «Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection,» *Journal Machine Learning Research*, vol. 13, pp. 27–66, 2012.
- [291] I. Guyon and A. Elisseeff, «An Introduction to Variable and Feature Selection,» *Journal of Machine Learning Research*, vol. 3, no. 7–8, pp. 1157–1182, 2003.
- [292] P. Y. Chen and P. M. Popovich, *Correlation: Parametric and nonparametric measures*, 137–139. Sage, 2002.
- [293] D. N. Reshef, Y. a. Reshef, H. K. Finucane, *et al.*, «Detecting Novel Associations in Large Data Sets,» *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.
- [294] D. Tjøstheim and K. O. Hufthammer, «Local Gaussian correlation: A new measure of dependence,» *Journal of Econometrics*, vol. 172, pp. 33–48, 2013.

- [295] R. Heller, Y. Heller, and M. Gorfine, «A consistent multivariate test of association based on ranks of distances,» vol. 100, no. 2, pp. 503–510, 2013.
- [296] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, «Measuring and testing dependence by correlation of distances,» *Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.
- [297] G. J. Székely and M. L. Rizzo, «The distance correlation t-test of independence in high dimension,» *Journal of Multivariate Analysis*, vol. 117, pp. 193–213, 2013.
- [298] G. J. Székely and M. L. Rizzo, «Partial distance correlation with methods for dissimilarities,» *Annals of Statistics*, vol. 42, no. 6, pp. 2382–2412, 2014.
- [299] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, «Speaker Verification Using Adapted Gaussian Mixture Models,» *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.
- [300] D. A. Reynolds, «Comparison of background normalization methods for text-independent speaker verification,» in *5th European Conference on Speech Communication and Technology (Eurospeech'97)*, Rhodes, Greece, 1997, pp. 963–966.
- [301] R. Vogt and S. Sridharan, «Explicit modelling of session variability for speaker verification,» *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.
- [302] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, «Speaker and Session Variability in GMM-Based Speaker Verification,» *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.
- [303] P. Kenny, P. Ouellet, and N. Dehak, «A study of interspeaker variability in speaker verification,» *Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.
- [304] 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.
- [305] J. Villalba and E. Lleida, «Handling i-vectors from different recording conditions using multi-channel simplified PLDA in speaker recognition,» in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 6763–6767.
- [306] D. Garcia-Romero and C. Y. Espy-Wilson, «Analysis of i-vector length normalization in speaker recognition systems,» *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 249–252, 2011.
- [307] R. Wallace and M. McLaren, «Total variability modelling for face verification,» *Iet Biometrics*, vol. 1, no. 4, pp. 188–199, 2012.
- [308] O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny, *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4057–4060, 2009.
- [309] N. Brümmer, L. Burget, J. H. Černocký, *et al.*, «Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006,» *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.
- [310] P. A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez, «Ordinal Regression Methods: Survey and Experimental Study,» *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016.

- [311] W.-Y. Deng, Q.-H. Zheng, S.-G. Lian, L. Chen, and X. Wang, «Ordinal extreme learning machine,» *Neurocomputing*, vol. 74, no. 1-3, pp. 447-456, 2010.
- [312] F. Stulp and O. Sigaud, «Many regression algorithms, one unified model: A review,» *Neural Networks*, vol. 69, pp. 60-79, Sep. 2015.
- [313] S. Calinon, «A tutorial on task-parameterized movement learning and retrieval,» *Intelligent Service Robotics*, vol. 9, no. 1, pp. 1-29, Jan. 2016.
- [314] J. S. Cardoso and J. F. Pinto, «Learning to Classify Ordinal Data : The Data Replication Method,» *Journal of Machine Learning Research*, vol. 8, pp. 1393-1429, 2007.
- [315] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, and P. Gutiérrez, «Metrics to guide a multi-objective evolutionary algorithm for ordinal classification,» *Neurocomputing*, vol. 135, pp. 21-31, Jul. 2014.
- [316] S. Baccianella, A. Esuli, and F. Sebastiani, «Evaluation Measures for Ordinal Regression,» in *2009 Ninth International Conference on Intelligent Systems Design and Applications*, IEEE, 2009, pp. 283-287.
- [317] T. Fawcett, «An introduction to ROC analysis,» *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.
- [318] T. Dunstone and N. Yager, Eds., *Biometric System and Data Analysis*. Boston, MA: Springer US, 2009.
- [319] A. P. Bradley, «The use of the area under the ROC curve in the evaluation of machine learning algorithms,» *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.
- [320] N. O'Brian, S. O'Brian, A. Packman, and M. Onslow, «Generalizability Theory I: Assessing Reliability of Observational Data in the Communication Sciences,» *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 3, pp. 711-717, 2003.
- [321] N. M. Webb, R. J. Shavelson, and E. H. Haertel, «Reliability Coefficients and Generalizability Theory,» in *Handbook of Statistics*, 06, vol. 26, 2006, pp. 81-124.
- [322] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, *et al.*, «Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program,» *European Archives of Oto-Rhino-Laryngology*, vol. 265, no. 4, pp. 465-476, Apr. 2008.
- [323] A. Moreno, D. Poch, A. Bonafonte, *et al.*, «Albayzin speech database: Design of the phonetic corpus,» in *EUROSPEECH*, 1993, pp. 175-179.
- [324] C. Draxler, «Introduction to the verbmobil-phondat database of spoken german,» in *Proc. 3rd Int. Conf. Practical Application Prolog*, 1995, pp. 201-212.
- [325] «Aplicación de las tecnologías de la información y comunicaciones» database. [Online]. Available: http://www.atlc.uma.es/index_atlc.html.
- [326] F. Schiel, «Automatic Phonetic Transcription of Non-Prompted Speech,» in *Proc. of the ICPhS*, San Francisco, Aug. 1999, pp. 607-610.
- [327] G. M. Weiss and F. Provost, «Learning when training data are costly: The effect of class distribution on tree induction,» *Journal of Artificial Intelligence Research*, vol. 19, pp. 315-354, 2003.
- [328] J. Mason and J. Thompson, «Gender effects in speaker recognition,» *Proc. ICSP-93, Beijing*, pp. 733-736, 1993.

- [329] J. M. Yentes, N. Hunt, K. K. Schmid, *et al.*, «The appropriate use of approximate entropy and sample entropy with short data sets.,» *Annals of biomedical engineering*, vol. 41, no. 2, pp. 349–65, Feb. 2013.
- [330] K. Chon, C. Scully, and S. Lu, «Approximate entropy for all signals,» *IEEE Engineering in Medicine and Biology Magazine*, vol. 28, no. 6, pp. 18–23, 2009.
- [331] A. Larcher, K. A. Lee, B. Ma, and H. Li, «Text-dependent speaker verification: Classifiers, databases and RSR2015,» *Speech Communication*, vol. 60, pp. 56–77, 2014.
- [332] D. Martínez González, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, «Voice pathology detection on the Saarbrücken Voice Database with calibration and fusion of scores using multifocal toolkit,» *Communications in Computer and Information Science*, vol. 328 CCIS, pp. 99–109, 2012.
- [333] D. Martínez González, E. Lleida, A. Ortega, and A. Miguel, «Score level versus audio level fusion for voice pathology detection on the Saarbrücken Voice Database,» *Communications in Computer and Information Science*, vol. 328 CCIS, pp. 110–120, 2012.
- [334] M. Sugiyama, *Introduction to statistical machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016.
- [335] D. Rossiter, «Statistical methods for accuracy assesment of classified thematic maps,» *Technical Note. Enschede: International Institute for Geo-information Science & Earth Observation (ITC)*, 2004.
- [336] J. Oates, «Auditory-Perceptual Evaluation of Disordered Voice Quality,» *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, Feb. 2009.
- [337] T. Ritchings, M. McGillion, and C. Moore, «Pathological voice quality assessment using artificial neural networks,» *Medical Engineering & Physics*, vol. 24, no. 7-8, pp. 561–564, Sep. 2002.
- [338] N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldan, «Automatic Assessment of Voice Quality According to the GRBAS Scale,» in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Aug. 2006, pp. 2478–2481.
- [339] R. Schönweiler, M. Hess, P. Wübbelt, and M. Ptak, «Novel approach to acoustical voice analysis using artificial neural networks,» *Journal of the Association for Research in Otolaryngology*, vol. 1, no. 4, pp. 270–282, 2000.
- [340] Z. Wang, P. Yu, N. Yan, L. Wang, and M. L. Ng, «Automatic Assessment of Pathological Voice Quality Using Multidimensional Acoustic Analysis Based on the GRBAS Scale,» *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 241–251, 2016.

ACKNOWLEDGEMENTS

This work was partially supported by the Ministry of Economy and Competitiveness of Spain under grants *DPI2017-83405-R1* and *TEC2012-38630-Co4-01*; and by Universidad Politécnica de Madrid, *Ayudas para la realización del doctorado* - RR01/2011.