UNIVERSIDAD POLITÉCNICA DE MADRID

**ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN**

**STUDY OF SPEECH AND CRANIOFACIAL
FEATURES IN OBSTRUCTIVE SLEEP APNEA
PATIENTS**

TESIS DOCTORAL

**FERNANDO MANUEL ESPINOZA CUADROS
INGENIERO DE TELECOMUNICACIÓN**

2018

**DEPARTAMENTO DE SEÑALES, SISTEMAS Y
RADIOCOMUNICACIONES**

**ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN**

**Study of Speech and Craniofacial features in
Obstructive Sleep Apnea Patients**

Autor: FERNANDO MANUEL ESPINOZA CUADROS

Ingeniero de Telecomunicación

Director: LUIS ALFONSO HERNÁNDEZ GÓMEZ

Doctor Ingeniero de Telecomunicación

Madrid, 2018

Colophon

This Thesis was typeset by the author using L^AT_EX2e. The main body of the text was set using a 11-points Computer Modern Roman font. The final postscript output was converted to Portable Document Format (PDF) and printed.

Copyright © 2018 by Fernando Manuel Espinoza Cuadros. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author. Universidad Politécnica de Madrid has several rights in order to reproduce and distribute electronically this document.

Department: Señales, Sistemas y Radiocomunicaciones
Escuela Técnica Superior de Ingenieros de Telecomunicación
Universidad Politécnica de Madrid (UPM)

PhD Thesis: **Study of Speech and Craniofacial features in
Obstructive Sleep Apnea Patients**

Author: **Fernando Manuel Espinoza Cuadros**
Ingeniero de Telecomunicación

Advisor: **Dr. Luis Alfonso Hernández Gómez**
Doctor Ingeniero de Telecomunicación

Year: 2018

Board named by the Rector of Universidad Politécnica de Madrid, on the ... of ... 2018

Board: **Dr. Eduardo López Gonzalo**
Universidad Politécnica de Madrid (UPM)

Dr. Rubén San Segundo Hernández
Universidad Politécnica de Madrid (UPM)

Dr. Ascensión Gallardo Antolín
Universidad Carlos III de Madrid (UC3M)

Dr. Doroteo Torre Toledano
Universidad Autónoma de Madrid (UAM)

Dr. Fares Alnajjar
University of Amsterdam (UvA)

The research described in this Thesis was developed at Grupo de Aplicaciones de Procesado de Señales (GAPS) between 2014 and 2018. This work was funded by the 2013 FPI scholarship from the Spanish Ministry of Economy and Competitiveness (MINECO) and the European Union (FEDER) as part of the TEC2012-37585-C02 (CMC-V2) project. The author are grateful to MD. José Daniel Alcazar Ramirez, and the Hospital Quirón Salud de Málaga for their support and assistance. The author also thanks to Sonia Martinez Díaz for her effort in collecting the OSA database that was used in this study.

A Teresa, mis padres y mi hermana.

Resumen

ESTA TESIS EXPLORA LA CARACTERIZACIÓN DE LA VOZ Y DEL FENOTIPO CRANEOFACIAL de pacientes diagnosticados con el síndrome de Apnea Hipopneas durante el Sueño (SAHS) mediante técnicas del estado del arte de tecnologías de caracterización de voz y procesamiento de imágenes para la caracterización de caras, así como el estudio y análisis de modelos supervisados de aprendizaje automático para la evaluación de las características de la voz y de las características craneofaciales como predictores del SAHS.

El síndrome de Apnea Hipopneas durante el Sueño (SAHS) es un tipo de trastorno respiratorio que afecta principalmente a los hombres de edad adulta, y se caracteriza por pausas recurrentes de la respiración durante el sueño debido al bloqueo total o parcial de las vías aéreas superiores. El diagnóstico del SAHS se realiza mediante la polisomnografía convencional. Para esta prueba es necesario que el paciente duerma en la unidad del sueño del hospital bajo la supervisión de un equipo médico con el propósito de registrar los patrones de respiración, ritmo cardíaco, y movimiento de las extremidades. Sin embargo, el procedimiento para el diagnóstico del SAHS es muy costoso debido al requerimiento de equipos y personal necesarios para la prueba, e invasivo para el paciente. Además, la lista de espera para el diagnóstico puede llegar a ser de un año.

Se han desarrollado muchos métodos como alternativa a esta problemática con el fin de reducir las listas de espera y acelerar la detección de los casos más severos. Entre estas pruebas se encuentran aquellas basadas en la detección de síntomas relacionados con el SAHS mediante cuestionarios realizados al paciente; inspección visual de la zona orofaríngea mediante el test de Mallampati, y análisis craneofacial mediante técnicas avanzadas de representación de imágenes tales como cefalometría, tomografía computarizada y análisis de imágenes por resonancia magnética. Entre las ventajas que pueden ofrecer estos métodos destacan la detección rápida de casos positivos y la priorización de casos severos de SAHS, así como su carácter menos invasivo en comparación con la polisomnografía convencional. No obstante, muchos de estos métodos son costosos y no son suficientemente generalizables.

Los primeros estudios para la evaluación del SAHS mediante técnicas de análisis de imágenes y caracterización antropométrica hallaron características anómalas en las estructuras de las vías aéreas de pacientes con SAHS. Por tanto, se pueden esperar patrones anómalos en el habla de los pacientes debido a la presencia de anomalías en las estructuras o funciones de sus vías respiratorias. Esta hipótesis fue confirmada por los primeros estudios basados en análisis acústico de grabaciones de pacientes diagnosticados con SAHS.

Los antecedentes expuestos anteriormente han llevado a proponer procedimientos menos costosos basados en el análisis de caras y grabaciones de voz de los pacientes para ayudar a la detección del SAHS, así como a la evaluación de su severidad. Por tanto, esta Tesis explora la caracterización del habla y el fenotipo craneofacial de pacientes diagnosticados con SAHS mediante técnicas de reconocimiento automático de locutor (i-vectors, supervectors) y caracterización de caras (características locales, modelado estadístico, características basadas en redes profundas).

Para las pruebas se empleó una base de datos de 729 pacientes (204 mujeres, 525 hombres), y las características de voz y craneofaciales se evaluaron mediante modelos supervisados de aprendizaje automático. Por otra parte, existen diferencias sobre cómo el SAHS afecta a mujeres y hombres, como por ejemplo los síntomas y factores de riesgo, los cuales pueden actuar como variables de confusión en el modelo para la detección del SAHS. Por tanto, es importante resaltar que los experimentos se realizaron para cada género y por separado.

Además, los primeros estudios para la detección del SAHS mediante el habla alcanzaron resultados favorables, sin embargo, después de un análisis de los mismos y de la metodología seguida, se encontraron muchas limitaciones, siendo algunas de estas: pocos datos de entrenamiento y el manejo incorrecto de los modelos de aprendizaje automático, provocando la aparición de falsos resultados.

Por tanto, la principal motivación que conduce al desarrollo de esta Tesis es la exploración de las técnicas de procesamiento automático del habla y caracterización automática de las caras, así como la evaluación de estas características mediante un modelo de validación exhaustiva con el objetivo de hacer frente a las limitaciones presentes en nuestra base de datos y evitar los típicos errores debido al manejo incorrecto de los modelos de aprendizaje automático. Por último, cabe destacar, de acuerdo a nuestro mejor conocimiento, que la presente Tesis es el único estudio que aborda la caracterización del fenotipo craneofacial y del habla en mujeres mediante el uso de procesamiento automático del habla y técnicas de caracterización facial.

Abstract

THIS THESIS EXPLORES THE SPEECH AND CRANIOFACIAL PHENOTYPE CHARACTERIZATION in Obstructive Sleep Apnea (OSA) patients by using the state-of-the-art speaker’s voice characterization technologies and image processing techniques for face recognition along with the study and analysis of supervised machine learning methods for evaluating these speech and craniofacial features as predictors of OSA severity.

The OSA is a common sleep-related breathing disorder affecting mainly men. It is characterized by recurring breathing pauses during sleep caused by a blockage of the upper airway (UA). The diagnosis of OSA is carried out at a sleep unit in a hospital by the polysomnography (PSG) test. This test requires an overnight stay of the patient at the sleep unit under the supervision of a clinician to monitor breathing patterns, heart rhythm, and limb movements, resulting in an invasive and costly method as well as the waiting list may exceed one year.

As an alternative to this test, many diagnosis schemes have been developed to help to reduce the waiting lists and accelerate the detection of severe cases such as questionnaires for OSA screening, and those based on medical-imaging, for instance oropharyngeal visual inspection (i.e. Mallampati test), and craniofacial assessment by means analysis techniques (e.g. cephalometry) of images created by advanced methods for visual representations (e.g. computed tomography, magnetic resonance imaging). Although these methods can help to increase the detection of positive cases as well as provide reliable results, most of them lack generalization such as questionnaires as well as they are costly and invasive for patients such as those used for craniofacial assessment.

Early studies for OSA assessment by using medical-imaging techniques and anthropometric characterization found out some evidence of abnormalities in upper airway structures in OSA subjects. Consequently, abnormal or particular speech features in OSA speakers may be expected from the altered structure or altered function of their upper airways.

These facts have led to proposing less costly procedures based on the analysis of patients’ facial images and voice recordings to help with OSA detection and severity assessment. There-

fore, this Thesis explores the speech and craniofacial characterization in Obstructive Sleep Apnea (OSA) patients by means of speech and craniofacial features based on automatic speaker recognition systems and face characterization techniques respectively: 1) supervectors and i-vectors, and 2) local features, statistical-model based features, and deep-learning-based features.

Using an existing database of 729 patients (204 women, 525 men), speech and craniofacial features were evaluated for OSA prediction by means supervised machine learning models. There are differences in how OSA affects men and women such as symptoms and risk factors, which could act as confounding factors. Therefore, it is important to emphasize that experiments were performed separately for each gender.

Furthermore, previous speech-based OSA detection studies have reached successful results, however, after a review of their results and methodologies, we found out several limitations, those being related to a small number of training samples as well as machine learning pitfalls in the methodology and validation scheme such as feature selection over a limited number of samples and high-dimensionality features resulting in a high probability of overfitting of the prediction model.

The ultimate motivation of this Thesis consists in exploring automatic speech processing and facial characterization techniques for OSA assessment on patients as well as their evaluation by means of an exhaustive validation scheme in order to face the limitations related to database size and to avoid the machine learning pitfalls due to the incorrect treatment of supervised learning models. Finally, to the best of our knowledge, the present Thesis is the unique study that approaches the speech and craniofacial phenotype characterization in women by using automatic speech processing and facial characterization techniques.

Table of Contents

Resumen	iii
Abstract	v
List of Figures	xiii
List of Tables	xvii
1 Introduction	1
1.1 Background	2
1.2 Motivation of the Thesis	4
1.3 Objectives	6
1.4 Outline	7
2 The Obstructive Sleep Apnea syndrome	9
2.1 Definition	9
2.1.1 Clinical profile and comorbidity	10
2.2 Risk factors and occurrence	12
2.2.1 Craniofacial anatomy	12
2.2.2 Excess body weight	12
2.2.3 Ethnicity, age, and gender	13
2.2.4 Occurrence	14
2.3 Clinical diagnosis	14
2.4 Physical and speech abnormalities in OSA	16
2.4.1 Craniofacial and upper airway abnormalities	17
2.4.2 Speech disorders	18
2.5 Summary	21

3	Databases and Methods	23
3.1	Introduction	23
3.2	Databases	24
3.2.1	Apnea Database v1.0	25
3.2.1.1	Data collection procedure	25
3.2.1.2	Speech Corpus	26
3.2.1.3	Speech and image data recording	29
3.2.2	Apnea Database v2.0	30
3.2.2.1	Data collection procedure	30
3.2.2.2	Speech and photographic data recording	31
3.2.3	Additional databases	34
3.3	Methods	37
3.3.1	Problem formulation	37
3.3.1.1	Regression approach	37
3.3.1.2	Classification approach	38
3.3.2	Supervised learning models	39
3.3.2.1	Support Vector Machine	39
3.3.2.2	Support Vector Regression	42
3.3.3	Model training and validation scheme	44
3.3.3.1	K -Fold Cross Validation and Grid search	45
3.3.4	Evaluation of performance	45
3.3.4.1	Evaluation in Regression approach	46
3.3.4.2	Evaluation in Classification approach	48
3.3.4.3	Agreement between two methods	54
3.3.4.4	Statistical analysis	55
3.3.5	Gaussian Mixture Models in Speaker Recognition	55
3.3.5.1	Definition	56
3.3.5.2	GMM training and MAP adaptation	57
3.4	Summary	58
4	Speech characterization in Obstructive Sleep Apnea	61
4.1	Introduction	61
4.2	Automatic voice characterization	62
4.2.1	Information in the speech signal	63

4.2.2	Session variability in the speech signal	64
4.3	Acoustic representation of OSA-related sounds	65
4.3.1	Mel-Frequency Cepstral Coefficients extraction	65
4.3.2	Utterance modeling	67
4.3.2.1	GMM approach	67
4.3.2.2	GMM-Supervector approach	69
4.3.2.3	I-vector approach	71
4.4	OSA assessment by speech features	73
4.4.1	Speech features validation	73
4.4.2	Apnea-Hypopnea Index prediction	75
4.4.2.1	Population without differences in clinical variables	82
4.4.2.2	The Apnea Databases' sentences	84
4.4.2.3	Session compensation analysis	84
4.4.3	OSA severity classification	86
4.4.3.1	OSA extreme cases classification	87
4.5	Discussion	88
5	Review of speech-based OSA detection models	93
5.1	Clinical variables as OSA predictors	93
5.2	OSA and speech connection	96
5.2.1	Differences in formants frequencies and bandwidths	96
5.2.2	Database size and patient's clinical condition	97
5.2.3	Confounder variables for OSA detection	98
5.2.4	Feature selection and gender-independent	99
5.3	Summary	101
6	Craniofacial phenotype characterization in Obstructive Sleep Apnea	103
6.1	Introduction	103
6.2	Review of Automatic Face Analysis	105
6.2.1	Face Alignment	107
6.3	Local features	108
6.3.1	Facial landmark detection	108
6.3.2	Uncalibrated measurements	110
6.3.2.1	Cervicomental contour area	110
6.3.2.2	Midface width	111

6.3.2.3	Tragion-ramus-stomion angle	112
6.3.2.4	Analysis of uncalibrated measurements	113
6.3.2.5	OSA severity classification	115
6.4	Statistical-model-based features	118
6.4.1	Active Appearance Model (AAM)	118
6.4.2	Face shape variation modeling	120
6.4.2.1	Cervicomental region	121
6.4.2.2	Midface-biocular region	121
6.4.2.3	Midface region	121
6.4.2.4	Analysis of shape-modeling-based features	122
6.4.2.5	OSA severity classification	123
6.5	Facial landmarks identification based on Deep Convolutional Network Cascade .	124
6.5.1	Performance evaluation	126
6.6	Deep learning-based features	127
6.6.1	Transfer Learning	129
6.6.2	VGG-Face-CNN as feature extractor	130
6.6.2.1	Problem formulation	130
6.6.3	Deep features validation	132
6.6.4	OSA assessment by deep features	132
6.6.4.1	Face surface regions employed by the classifier	134
6.6.4.2	Facial features analysis in OSA prediction	135
6.7	Summary	137

7 Obstructive Sleep Apnea Syndrome in Women 139

7.1	Introduction	139
7.2	Clinical variables as OSA predictors	141
7.3	OSA assessment by speech features	144
7.3.1	Apnea-Hypopnea Index prediction	144
7.3.2	OSA severity classification	146
7.4	Craniofacial phenotype characterization in OSA women	149
7.4.1	Uncalibrated measurements	149
7.4.2	Deep learning-based features	151
7.4.2.1	Features validation	152
7.4.2.2	OSA assessment by deep features	152

7.4.2.3	Analysis of facial surface elements for OSA assessment	153
7.5	Summary	155
8	Conclusions, Research Contributions and Future Work	157
8.1	Summary of the Thesis	157
8.2	Conclusions	159
8.3	Research Contributions	161
8.4	Future work	163
	Bibliography	165

List of Figures

2.1	Block diagram describing the Apnea-Hypopnea episode.	11
3.1	Sentences of Apnea Databases v1.0 and v2.0, including IPA phonetic transcription and English translation.	28
3.2	Frontal and profile photographic samples from the Apnea Database v2.0.	34
3.3	Block diagram describing Apnea-Hypopnea Index (AHI) and clinical variables estimation.	38
3.4	Block diagram describing OSA severity assessment by means of a classification model and using speech features.	39
3.5	Representation of the SVM underlying idea and representation of a separable data classification and basic elements of the SVM.	41
3.6	Representation of the SVR underlying idea and the soft margin loss setting for a linear SVR.	43
3.7	Block diagram describing k-fold cross-validation and grid-search process for SVM/SVR model.	46
3.8	Example of probability density function of scores resulting from binary classification.	49
3.9	Description of the optimal values of sensibility, specificity, and threshold estimation in the ROC curve.	52
3.10	Bland-Altman plot for average PEFr by two methods.	53
3.11	Representation of components of a GMM	56
3.12	A Maximum Posteriori Adaptation process representation.	58
4.1	Mel-frequency coefficient extraction procedure.	66
4.2	Modular representation of training and testing phases performed in speech-based OSA detection studies by means of the GMM approach.	68
4.3	GMM-Supervector modeling.	69

4.4	a) Evolution of the MAE for male speakers' AHI prediction in function of the numbers of i-vectors considered. b) Eigenvalues associated with the i-vectors estimated sorted in descending order.	78
4.5	a) Plot regression for male speakers' AHI estimation by means i-vectors of dimension 100. b) Bland-Altman plot for male speakers' AHI estimation results by means i-vectors of dimension 100. Points inside the rectangular area (black line) are the percentage of diagnostic agreement between the two methods.	79
5.1	a) Plot regression for male speakers' AHI estimation using clinical variables. b) Bland-Altman plot for male speakers' AHI estimation results using clinical variables.	95
6.1	Block diagram summarizing basic approach of automatic face recognition system.	105
6.2	Landmarks on frontal and profile view extracted using the <code>aam_tools</code>	109
6.3	Block diagram describing the landmarking process by model fitting approach. . .	109
6.4	Measurements used for the cervicomenta contour area (left) and measurements used for midface width (right).	111
6.5	Trigion-ramus-stomion angle (left) and lateral cephalometric images of normal and OSA subjects (right).	112
6.6	Block diagram describing the OSA severity classification using uncalibrated craniofacial photographic measurements.	113
6.7	Cervicomenta region.	121
6.8	Midface-biocular region (left) and midface region (right).	122
6.9	Effects of varying the first shape parameter of the cervicomenta region.	124
6.10	Landmarks on frontal view extracted by Face++ Landmarks API.	126
6.11	Scheme of VGG-Face CNN descriptor used as feature extractor.	131
6.12	ROC for OSA detection by using uncalibrated measurements and deep features. .	133
6.13	Heat maps showing the degree to which masking a given part of an image changes the performance for the AHI and clinical prediction in the male population. . . .	135
6.14	Composite men faces built by averaging faces classified as OSA and Control by using the deep features.	136
6.15	Average facial landmarks on frontal view extracted from male population' facial photographs and using Face++ Landmarks API.	137
7.1	Regression and Bland-Altman plots for female a) and male b) speakers' AHI estimation by using clinical variables.	145

7.2	a) Regression and Bland-Altman plots for female speakers' AHI estimation by using i-vectors. b) Regression and Bland-Altman plots for male speakers' AHI estimation by using i-vectors	147
7.3	ROC for OSA detection in men and women by using i-vectors.	148
7.4	ROC for OSA detection by using uncalibrated measurements and deep features. .	153
7.5	Heat maps showing the degree to which masking a given part of an image changes the performance for the AHI and clinical prediction in the female population. . .	154
7.6	Composite women faces built by averaging faces classified as OSA and Control by using the deep features.	155

List of Tables

2.1	AHI thresholds and degree of daytime sleepiness for OSA severity assessment defined by American Academy of Sleep Medicine.	15
3.1	Summary on the male subjects, including the mean and standard deviation ($\pm$ SD) for each of the clinical variables collected along with speech and image records .	32
3.2	Summary on the female subjects, including the mean and standard deviation ($\pm$ SD) for each of the clinical variables collected along with speech and image records.	32
3.3	Number of utterances for each type of sentence in Apnea Database v2.0.	33
3.4	Number of utterances for Control and OSA groups in male population in terms of each type of sentence.	33
3.5	Number of utterances for Control and OSA groups in female population in terms of each type of sentence.	33
3.6	Description of databases included in the development dataset	36
3.7	Description of Grid search for selecting the optimal hyperparameters for a classification model.	47
3.8	Confusion matrix for OSA detection test	50
3.9	Description of diagnostic agreement approach for determination of diagnostic errors between PSG and test methods.	55
4.1	Description of parameters used by functions in order to extract the speech features (MFCCs, Supervector, i-vector) and train SVM and SVR models for OSA classification and predict AHI.	70
4.2	Performance comparison among prediction models for male speakers' age estimation	74
4.3	Performance comparison among prediction models for male speakers' height estimation.	74

4.4	Performance comparison among prediction models for male speakers' weight estimation.	75
4.5	Male speakers' AHI and clinical variables estimation by the prior estimator. . . .	76
4.6	Male speakers' AHI and clinical variables estimation using supervectors and SVR (linear kernel).	76
4.7	Male speakers' AHI and clinical variables estimation using i-vectors and SVR (linear kernel).	77
4.8	Male speakers' AHI and clinical variables estimation using i-vector and SVR (RBF kernel).	77
4.9	Percentage of estimated and overestimated AHI values of the non-OSA (AHI < 10) male population using i-vectors and supervectors.	79
4.10	Percentage of estimated, underestimated and overestimated AHI values of the mild-OSA ($10 \leq \text{AHI} \leq 30$) male population using i-vectors and supervectors. . .	80
4.11	Percentage of estimated and underestimated AHI values of the severe OSA (AHI > 30) male population using i-vectors and supervectors.	80
4.12	Diagnostic agreement approach for male speakers' AHI estimation using i-vectors and supervectors.	81
4.13	Diagnostic Agreement percentage for each OSA severity group using i-vectors of dimension 100, and supervectors.	81
4.14	Contrast analysis among Control (AHI < 10) and OSA (AHI ≥ 10) groups on a subject without differences on body mass index (BMI) and age.	82
4.15	Number of utterances for control and OSA groups in uniform male population in terms of each type of sentence.	83
4.16	AHI prediction in a male population without differences in clinical variables by using i-vectors of dimension 100.	83
4.17	Male speakers' AHI estimation using i-vectors of dimension 100 for each type of sentence.	84
4.18	OSA severity classification for male speakers using the SVM model (linear kernel). .	86
4.19	OSA severity classification for male speakers using the SVM model (RBF kernel). .	87
4.20	Contrast among Control (AHI < 10) and severe OSA (AHI > 30) groups including mean and standard deviation (mean $\pm$ SD).	88
4.21	OSA extreme and in-between cases classification for male speakers using SVM model (linear kernel) and i-vectors of dimension 100	88

4.22	Male speakers' AHI estimation using i-vectors and SVR (linear kernel). The development data used for the Total Variability subspace training is the Apnea Database v2.0.	90
4.23	OSA severity classification for male speakers using SVM model (linear kernel). The development data used for the Total Variability subspace training is the Apnea Database v2.0.	91
5.1	Pearson's correlation between clinical variables and AHI on the male population.	94
5.2	Male speakers' AHI estimation using clinical variables.	94
5.3	OSA classification severity for male speakers using clinical variables.	94
5.4	Test characteristics of previous research using speech analysis and supervised models for OSA severity classification and AHI prediction.	96
5.5	Wilcoxon two sampled test for MEAN_HNR_VA_A contrasting gender and group of extreme OSA male speakers.	100
5.6	Speaker's AHI estimation using supervectors generated by five high-order cepstral and LPC coefficients.	101
6.1	Contrast analysis between Control and OSA groups for uncalibrated craniofacial measurements on the male population, including the mean and standard deviation ($\pm SD$).	113
6.2	Contrast analysis between Control ($AHI < 10$) and OSA ($AHI \geq 10$) groups in the male population for uncalibrated craniofacial measurements on a subset without differences on body mass index and age.	114
6.3	Pearson's correlation between uncalibrated craniofacial measurement, clinical variables, and AHI on the male population.	115
6.4	OSA severity classification for male patients using clinical variables and uncalibrated craniofacial photographic measurements.	116
6.5	Percentage of deviation of shape parameters respect to the total deviation of shape parameters from <i>midface-biocular region</i>	122
6.6	Pearson's correlation between shape parameters, clinical variables, and AHI on the male population.	123
6.7	OSA severity classification for male patients by using shape parameters, and fusion at feature level with clinical variables and uncalibrated craniofacial photographic measurements.	124

6.8	Pearson's correlation between semi-automatically (aam_tools + human-supervised stage) and automatically (Face++) determined uncalibrated measurement midface width, clinical variables and AHI.	126
6.9	Classification results for OSA prediction by semi-automatic method (i.e. aam_tools) and automatic method (Face++) and using shape parameters from midface and midface-biocular regions.	127
6.10	Clinical variables estimation for male patients using the deep features extracted from VGG-Face.	132
6.11	OSA severity classification results for male patients by using deep features extracted from later layers of VGG-Face.	133
6.12	Correlation values between landmarks' distances from mandibular region, BMI, Cervical perimeter and the AHI.	137
7.1	Summary on the women and men, including the mean and standard deviation (mean $\pm$ SD) for each of the clinical variables and AHI. The last column includes <i>p</i> -values estimated by non-parametric Wilcoxon rank-sum test applied to each variable considering only women and men.	140
7.2	Pearson's correlation between clinical variables and AHI on the female population.	141
7.3	Female subjects' AHI estimation by using clinical variables.	142
7.4	OSA classification severity for female subjects by using clinical variables.	142
7.5	Diagnostic agreement approach for female and male subjects' AHI estimation by using clinical variables.	143
7.6	Multiple linear regression model for AHI prediction in women. The model is fitted to the training data by means stepwise method.	143
7.7	Multiple linear regression model for AHI prediction in men. The model is fitted to the training data by means stepwise method.	144
7.8	Female speakers' AHI and clinical variables estimation by using i-vectors and SVR-linear kernel.	146
7.9	Diagnostic agreement approach for female and male speakers' AHI estimation by using i-vectors of dimension 200 for women and 100 for men, and SVR-linear kernel	146
7.10	OSA severity classification for female speakers by using i-vectors and SVM model (linear kernel).	146
7.11	AHI value prediction in a female and male population without differences in clinical variables by using i-vectors.	148

7.12	Contrast analysis between Control ($AHI < 10$) and OSA ($AHI \geq 10$) groups for uncalibrated craniofacial measurements on the female population, including the mean and standard deviation ($\pm SD$).	149
7.13	Pearson's correlation between uncalibrated craniofacial measurement, clinical variables, and AHI on the female population.	150
7.14	Contrast analysis between Control ($AHI < 10$) and OSA ($AHI \geq 10$) groups in female population for uncalibrated craniofacial measurements on a subset without differences on body mass index and age.	151
7.15	OSA severity classification for boths genders by using uncalibrated craniofacial photographic measurements.	151
7.16	Clinical variables estimation for female patients by using the deep features extracted from the pre-trained net VGG-Face.	152
7.17	OSA severity classification for female and male patients by using deep features extracted from the full-connected ₆ (Layer 14) of the VGG-Face CNNs.	152

Chapter 1

Introduction

IN THE LAST DECADE, sleep disorders are receiving increased attention as a cause of daytime sleepiness ([Dinges, 1995](#)), impaired work, and traffic accidents ([Horne and Reyner, 1995](#)) and are associated with hypertension, heart failure, arrhythmia, and diabetes ([Buxton and Marcelli, 2010](#)).

The Sleep Apnea is a type of sleep-related breathing disorder, and it is characterized by the interruption of breathing during sleep with episodes duration greater than 10 seconds that may repeat more than 30 times in an hour ([Kushida, 2008](#)). Considering that there are many factors that may influence the sleep apnea such as different causes of breath cessation, three types of sleep apnea are defined in the OSA-epidemiology studies. The most common form is the Obstructive Sleep Apnea ([Lam et al., 2010](#)), which is highly prevalent in adult male population (30-60 years). The cessation of breath is because of the obstruction in the upper airway at the level of pharynx due to the lack of muscular tone, tissue softness or excess of regional adipose tissue. In contrast, the breathing disruption regularly during sleep due to concomitant instability in the body's feedback that controls the respiration is defined as Central Sleep Apnea ([Kushida, 2008](#)). The combination of both types is defined as Mixed Sleep Apnea.

With the objective of detecting and quantifying the apnea episodes, patients are required to overnight at sleep disorders unit of hospital plugged to many devices to monitor their life patterns such as breath effort, cardiac evolution, outgoing air pressure, etc. Despite the high accuracy for apnea detection, this method introduces costly procedures to apnea diagnosis and more invasive procedure for patients. Furthermore, the waiting list could last up more than one year, such as in Spain ([Puertas et al., 2005](#)). In addition, the lack of diagnostic equipments at sleep units (i.e. PSG), the high costs, and inconveniences of diagnosis protocols have contributed to there is still a high number of undiagnosed cases ([Puertas et al., 2005](#)). Consequently, these limitations along

with the sleep apnea relation to impaired work performance and traffic accidents ([Lloberes et al., 2000](#)), have brought clinicians attention to outline alternative diagnostic methods or screening tools in order to provide reliable results on the presence or absence of sleep apnea symptoms.

Many screening methods have been developed, which have reached high rates for apnea detection, for instance, the use of questionnaires for sleep apnea screening and consist in asking a set of questions related to clinical symptoms of patients ([Chung et al., 2008](#)) or oropharyngeal visual inspection to assess the grade of oropharyngeal appearance used to assess difficulty of intubation, thus the collapsibility grade of upper airway in OSA patients ([Mallampati et al., 1985](#)). Nevertheless, despite the good results reported by these methods, they must be carefully managed due to the poor generalization provided by their methodology.

Recent advances in speech and image processing technologies have prompted their use to predict some diseases or syndromes. The main advantage relies on the fact that they may be used to evaluate patient's condition in a quick, non-invasive, cost-effective and more accurate way than other methods. In the case of speech-related apnea detection, previous studies reported positive results by using speech processing technologies for speech abnormalities detection related to sleep disorders such as the Obstructive Sleep Apnea (OSA) syndrome ([Fernández-Pozo et al., 2009](#); [Zigel et al., 2008](#)), and speech disorders related to neurodegenerative diseases: Parkinson's disease ([Tsanas et al., 2011](#)), Moderate Cognitive Impairment ([Espinoza-Cuadros et al., 2014](#)), and Alzheimer's disease ([Gómez-Vilda et al., 2011](#)). In the same manner, the use of image processing techniques have increased due to potential to characterize physical traits associated with the symptomatic population. For instance, the craniofacial measurements analysis for OSA assessment ([Guilleminault et al., 1984](#); [Lowe et al., 1995](#); [Schwab et al., 2003](#)), and facial-analysis-based syndromes diagnosis ([El-Rai et al., 2015](#); [Kitano et al., 1997](#); [Todd et al., 2006](#)).

1.1 Background

Among sleep disorders, the Obstructive Sleep Apnea (OSA) syndrome is one of the most common forms of sleep-related breathing disorders ([Lam et al., 2010](#)). The OSA prevalence is 10% for adult men and 3% for adult women between ages of 30 and 49 years old and 17% for adult men and 9% for adult women between the ages of 50 and 70 years ([Peppard et al., 2013](#)).

The Obstructive Sleep Apnea (OSA) is characterized by frequent episodes of breathing pauses caused by partial or total occlusion in the upper airway at the level of the pharynx. The detection of OSA in patients is comprised of two stages: pre-screening and diagnosis. First, a visual screening of patients is required. By means this test, the clinician assesses the symp-

toms described by patients such as excessive daytime sleepiness, loud snoring at night, morning headaches, fatigue, apnea episodes witnessed by a bed partner, etc. Based on test results, the patient will be discarded or not for next stage; the diagnosis by polysomnography (PSG). The polysomnography is the gold standard diagnostic test for OSA ([American Association of Respiratory Care-Association of Polysomnography, 1995](#)). This test requires an overnight stay of the patient at the sleep disorders unit within a hospital to monitor breathing patterns, heart rhythm, and limb movements. As a result of this test, the OSA is assessed by the Apnea-Hypopnea Index (AHI), which is computed as the average number of apnea and hypopnea episodes (partial and total breath cessation episodes respectively) per hour of sleep. Due to its high reliability, this index is used to describe the severity of patient's condition. According to American Academy of Sleep disorder, an AHI of 15 sets the threshold from mild to moderate OSA. A detailed description of these values is given in section 2.4. Despite the high reliability of PSG test, its use is limited to the availability of a laboratory-based polysomnography (PSG) on the sleep disorders unit of the hospital, for instance in Spain there are 0.49 sleep disorders units equipped per 100.000 population ([Durán-Cantolla et al., 2004](#)). Consequently, this lack of resources increases the cost of OSA diagnosis, and the number of undiagnosed cases. These facts have motivated the development of other OSA screening methods, such as the home apnea diagnosis devices ([Flemons et al., 2003](#)), indirect identification of OSA from oxygen-saturation ([Álvarez et al., 2008](#)), snoring sound ([Solà-Soler et al., 2012](#)), and nocturnal-breathing patterns [Solà-Soler et al. \(2008\)](#) analysis.

The breath cessation during sleep in patients suffering from OSA due to the blockage of airflow pathway along with the evolutionary changes in the acquisition of speech ([Davidson, 2003](#)), prompted to the hypothesis to the appearance of OSA due to a characteristic configuration of the upper airway (UA), consequently, the appearance of abnormal patterns during the speech production in patients. On the basis of that hypothesis, many speech-based OSA detection approaches have been conducted, ranging from first studies based on the perceptual analysis ([Monoson and Fox, 1987](#)), to the speech processing techniques ([Fernández-Pozo et al., 2009](#); [Zigel et al., 2008](#)). The upper airway (UA) configuration of OSA subjects is connected to the existence of physical abnormalities, in particular, those related to craniofacial morphology, where most of those relationships were found by means of image-analysis techniques as cephalometry ([Guilleminault et al., 1984](#)), computed tomography ([Lowe et al., 1995](#)), and magnetic resonance imaging ([Schwab et al., 2003](#)). Nevertheless, despite the detailed description of craniofacial morphology provided by those methods, most of them are limited to research applications due to high-cost analysis procedures as well as being more invasive and risky procedure for patients.

On the basis of the existence of craniofacial abnormalities in OSA patients, as described above, [Lee et al. \(2009a\)](#) considered the hypothesis of the existence of craniofacial morphological phenotype in OSA-diagnosed patients. Under this hypothesis, [Lee et al. \(2009a\)](#) brought out a novel quantitative photographic approach for craniofacial phenotyping characterization in OSA subjects based on the anthropometry and the photogrammetry, providing new insights into the craniofacial morphological phenotyping in OSA population. As compared to previous image-analysis techniques such as computed tomography, this new approach is non-invasive, less costly, and readily accessible

1.2 Motivation of the Thesis

The research work presented in this Dissertation is part of an ongoing project developed by the research group Signal Processing Application group (GAPS) from the Universidad Politécnica de Madrid (UPM) since 2008. The project started with two ultimate goals: first, the development of a database of speech recordings and facial photographs of subjects diagnosed with Obstructive Sleep Apnea. Second, the study of the connection between speech and the Obstructive Sleep Apnea. Under this project, early studies were carried out providing important results such as new insights of possible evidence on the connection between speech and OSA based on the acoustic analysis on the speech recordings by means of speech processing techniques ([Fernández-Pozo et al., 2009](#); [Blanco-Murillo et al., 2011](#); [Montero-Benavides et al., 2016](#)). Later, with more photographs samples in the database and considering the breaking-through advances in diagnosis syndromes based on facial patterns, an extra goal was added to the project. The third goal is the search for facial patterns on photographs to extract information related to OSA subjects. Therefore, the work presented in this Thesis seeks to contribute to the search for acoustic and facial patterns for characterizing the OSA population.

Bearing in mind the good performance obtained by speech and craniofacial features in the state-of-the-art OSA recognition, as described in previous studies, the ultimate motivation of this Thesis is clear: exploring the most discriminative state-of-the-art speaker recognition and automatic facial characterization techniques for extracting inherent traits from OSA patients. Nevertheless, additional aspects and observations from last advances in the state-of-the-art speech-and-facial-based OSA detection have also motivated the work conducted in this Dissertation. In order to provide clarity to the Dissertation's motivation, aspects and observations are grouped into four scenarios, which are the starting points of the research work of this Dissertation.

- **Scenario 1:** besides the subject's identity information embedded into speech (e.g. voice

quality, prosodic features), there are other types of information that also contribute to the recognition task. Among them, there are physical attributes such as age and height, which can be successfully estimated by speech (Poorjam et al., 2015a) and facial features (Kwon and Lobo, 1994). Considering the hypothesis of the existence of connection between the speech and OSA due to the abnormalities in the upper airway structure configuration in OSA patients (Fox et al., 1989; Robb et al., 1997; Solé-Casals et al., 2014), along with the good results reported by previous studies for OSA prediction by using automatic speech processing techniques Fernández-Pozo (2011); Zigel et al. (2008), this Thesis explores the state-of-the-art speaker recognition technique to extract information related to speech abnormalities patterns and predict OSA in subjects.

- **Scenario 2:** despite good performances reported in previous speech-based OSA detection studies (Fernández-Pozo, 2011; Zigel et al., 2008; Solé-Casals et al., 2014), there is some evidence that most of those good results could be related to the incorrect treatment of some machine-learning related pitfalls. For instance, physical attributes such as weight, age, and body mass index can be accurately predicted by speech features. Nevertheless, most of these attributes are related to OSA risk factors, where some of them could be correlated with OSA. Consequently, the prediction model for predicting OSA by using speech features could be influenced by these attributes as they could act as confounding factors that could lead to false discoveries (Foster et al., 2014). Another case study, the use of feature selection on high dimensionality features or incorrect validation methods for prediction model (Smialowski et al., 2010), which could prompt overfitting. Considering these points, besides exploring speech and facial features for OSA prediction and considering the nature of the dataset used in the experiments, the Thesis carried out exhaustively the treatment of machine learning pitfalls and validation process for prediction models, as well as a review of the methodology applied by previous studies in the state-of-the-art speech-based OSA recognition.
- **Scenario 3:** on the basis of (Lee et al., 2009a, 2010), the characterization of facial phenotyping in OSA subjects by craniofacial measurements extracted from facial landmarks, previously annotated accurately on face photographs, provides discriminative information related to abnormalities in the upper airway structure. Considering this finding, we aim for testing similar craniofacial measurements but adapted to our uncontrolled photography capture process. In the same manner, considering the break-through advances in computer vision field, in particular, deep-learning-based facial characterization, this Dis-

sertation aims for testing these advanced techniques in order to represent information related to abnormal facial patterns to predict OSA in subjects.

- **Scenario 4:** gender-related features can become strong confounding factors that could lead to false discoveries. According to epidemiology studies of OSA, there are differences in how the OSA-related symptoms affect female and male. In the same manner, there are also differences in voice and facial characteristics between women and men. Therefore, an important aspect of this Thesis is the exploration of speech and facial patterns in the female population, which has not been exhaustively considered in the state-of-the-art speech-based OSA detection and OSA craniofacial morphological characterization.

In summary, the approach of critical aspects, as described above, is required to review exhaustively the methodology applied by previous studies. Furthermore, the approach of advanced speech and facial characterization techniques is required to find out features that may carry information associated with OSA, and provide a high robustness against different sources of variability.

1.3 Objectives

A brief description of the Dissertation's goal is testing whether the state-of-the-art speaker and automatic facial recognition techniques can be effective to predict OSA in subjects. Nevertheless, the foundations of this Dissertation rely on four prime objectives, as described in the following:

- Considering both the high robustness of the state-of-the-art speaker recognition techniques for representing speaker's inherent information, and the connection between speech disorders and the upper airway structure configuration in OSA subjects, the first objective is the exploration of the Supervectors and I-vectors to extract information related to speech disorders in OSA subjects.
- Review of previous studies for speech-based OSA detection, which reported successful results for OSA detection. in particular results and the experimental framework in order to search for machine-learning related pitfalls due to the incorrect address of some machine-learning related pitfalls such as feature selection approach on low dimensionality feature space, and the incorrect address of validation methods when working with features as compared to a number of cases.

- On the basis of the successful representation of craniofacial traits of OSA patients by surface facial measurements extracted from facial photographs, following a high controlled photographic capture and facial landmarking procedure (Lee et al., 2009a), this Thesis explores the use of surface facial measurements but adapted to a less controlled photographic capture and facial landmarking procedure. Two types of craniofacial representation are approached: local features (i.e. uncalibrated measurements and shape- and appearance-modeling based features) and deep-learning based features.
- The differences in OSA prevalence between women and men, explain the different factors to determine the OSA pathogenesis. Factors such as aging, central obesity are common in both genders, nevertheless, speech and craniofacial differences between both genders contribute to differences in OSA pathogenesis between women and men. In addition, the hormonal factor also plays an important role in the appearance of OSA in women. Consequently, these factors might contribute to differences in speech and craniofacial characterization of women as compared to men. Therefore, this Dissertation analyzes the differences in speech and facial features between women and men as well as identification of such factors that contribute to these differences.

1.4 Outline

The outline of this Dissertation has been structured in eight chapters. Essentially, chapters are structured as follows:

- **Chapter 1** introduces a brief description of sleep disorders, emphasizing the obstructive sleep apnea. This chapter also describes the background required to approach the Thesis, the motivation and the objectives of this Dissertation.
- **Chapter 2** briefly provides a general description of the epidemiology of the Obstructive Sleep Apnea (OSA), including the pathogenesis, risk factors, prevalence, and diagnosis. This chapter also reviews the connection between speech, physical abnormalities in the upper airway's structure and OSA.
- **Chapter 3** presents the database designed and collected for the GAPS project in collaboration with the Hospital Clínico de Málaga (Apnea Database v1.0) and the Hospital Quirón Salud de Málaga (Apnea Database v2.0). It also presents the collection data procedure (protocol, design procedure, inclusion criteria). This chapter also presents the methodology and the experimental framework followed, which includes a description of the problem

formulation based on the goals, a brief description of supervised machine learning methods used for classification and regression, and the validation schemes for the experiments.

- **Chapter 4** presents the study of connection between speech and OSA. This chapter describes the state-of-the-art speech characterization techniques along with their analysis and evaluation for OSA assessment through supervised learning models.
- **Chapter 5** reviews previous studies performed in the state-of-the-art speech-based OSA detection. This chapter approaches the discussion of possible evidence related to machine-learning pitfalls such as feature selection approach on high dimensionality feature space and validation methods.
- **Chapter 6** presents the study of the craniofacial characterization in OSA patients by means of image-processing techniques: local image descriptors and deep-learning based techniques.
- **Chapter 7** presents the study of speech and craniofacial features for OSA prediction in women. This chapter also contrasts the obtained results with those obtained in the male population.
- **Chapter 8** concludes the Dissertation summarizing the main results and outlining future research lines.

Chapter 2

The Obstructive Sleep Apnea syndrome

THIS CHAPTER PRESENTS A REVIEW OF THE OBSTRUCTIVE SLEEP APNEA (OSA) syndrome. The review includes a general scope about the grounds of the OSA such as the pathogenesis and consequences on patient's health condition, as well as a review of the techniques currently used in its diagnosis. This chapter also reviews the connection between OSA and the appearance of abnormalities in the OSA patients' upper airway structure, consequently, the appearance of abnormal speech and craniofacial traits. Therefore, the thesis reviews the basis of these connections along with previous studies for speech-and-image-based OSA detection.

2.1 Definition

The Obstructive Sleep Apnea (OSA) syndrome is a sleep-related breathing disorder. Early epidemiology studies reported prevalences of 3-7% of male adults and 2-5% of female adults between 30 and 70 years ([Young et al., 1993](#)). In the last two decades, there has been an increase of the OSA prevalence in general population, as reported by [Peppard et al. \(2013\)](#): a prevalence of 10% for men and 3% for women between 30-49 years old and 17% for men and 9% for women between 50-70 years old. Moreover, the OSA prevalence is associated with central obesity, an excess of regional adipose tissue and craniofacial abnormalities ([Cakirer et al., 2001](#); [Schellenberg et al., 2000](#); [Stradling and Crosby, 1991](#)).

The OSA is characterized by repetitive, partial (hypopnea) or complete (apnea) blockage of the upper airway (UA) during sleep at the level of the pharynx, with breath's cessation episodes greater than 10 s. at a time. This time-period was first established by [Guilleminault et al.](#)

(1976) and adopted by consensus, but in the community there are some disagreements with this time-period since it does not regard other aspects associated with comorbidities such as heart failure or breathing disorders, aging, and sex gender. Nowadays, besides the established time-period for breath cessation, an apnea episode is defined as a reduction in airflow of 90% (Puertas et al., 2005), and hypopnea episode is defined as a reduction in airflow of 30-50% with a decrease in oxyhaemoglobin (Lam et al., 2010; Punjabi, 2008).

The blockage of the UA at the level of the pharynx occurs when the normal reduction in pharyngeal dilator muscle tone on the onset of sleep is superimposed on a narrowed and/or highly compliant pharynx. This episode starts with the onset of sleep when the patient has fallen sleep the upper airway (UA) resistance increases, consequently, the volume of lung decreases due to the airflow reduction (Heinzer et al., 2006). To further complicate, the lack of stiffness of UA's soft tissues or fat deposition surrounding, the pharyngeal cavity elicits a complete (apnea) or partial (hypopnea) reduction of the airflow which causes a decrease of oxygen in the blood. Although there is no a consensus about the definition of reduction of airflow, which defines the grade of obstruction, this Thesis followed the definition established in (Puertas et al., 2005; Lam et al., 2010; Punjabi, 2008): the complete reduction of the airflow (apnea) is defined as a reduction greater than 90% of airflows and the partial reduction of the airflow (hypopnea) is defined as a reduction of 30-50% of airflow followed by arousal from sleep or a decrease in oxyhaemoglobin saturation of 3-4%. So as to overcome airflow cessation, balance gases concentrations and recover normal breath, the brain sends a signal to reopen the UA and restore the normal ventilation, returning the patient to the wakeful state. This last stage is referred to as micro-awake or arousal. At this point, the patient falls asleep and once again the apnea-hypopnea episodes repeat. A representation of this episode is depicted in Figure 2.1. Besides the pharynx's anatomical configuration enclosed along its length by bones, nasal turbinates, hard palate of the maxilla, mandible, hyoid bone and soft tissues as soft palate, tonsillar pillars, pharyngeal mucosa and muscles epiglottis, an excess of tissues mass and/or presence of morphological abnormalities on pharynx airway is also common in OSA patients (Ryan and Bradley, 2005).

2.1.1 Clinical profile and comorbidity

According to (Puertas et al., 2005), the typical clinical profile of OSA subjects is characterized by chronic snoring (loud and disruptive), excessive daytime sleepiness, and obesity, which is considered a high OSA risk factor, but it is also linked to daytime sleepiness (Vgontzas et al., 1998). Moreover, other symptoms have also been reported in most of the diagnosed cases

such as insomnia, pauses in breathing with gasping and snorting episodes. Previous studies reported a strong relationship between OSA, daytime sleepiness, and impaired cognitive function, which have a negative impact on OSA subjects life quality (Broughton et al., 1978), and may contribute to impaired work and traffic accidents (Horne and Reyner, 1995). The presence of OSA in subjects is connected to the appearance of causes of mortality in adults such as hypertension, cardiovascular, and cerebrovascular diseases. (Young et al., 2002b). For instance, as regards hypertension, the episodes of apnea and hypopnea during sleep produce an acute increase of blood pressure and nocturnal episodes of hypoxia and arousal due to upper airway's blockage may lead to sustained elevation of blood pressure via the pathophysiologic mechanism. But, this cause-effect is not reversible as the potential for remediating hypertension by treating OSA remains unclear (Punjabi, 2008). In the same manner, the obstructive respiratory events may lead disturbances in cardiovascular function including hearth hypertrophy (Guidry et al., 2001), heart failure (Bradley, 1992), and plaque ruptures and subsequent cardiovascular or cerebrovascular events (Hedner et al., 1994).

Figure 2.1: Block diagram describing the Apnea-Hypopnea episode.

Besides the comorbidities associated with OSA, there also exist risk factors which play a role in the OSA predisposition and may cause an acute increase in comorbidities severity. Therefore, the diagnosis and treatment must be correctly conducted to decrease the higher number of undiagnosed cases as well as the OSA-related comorbidities.

2.2 Risk factors and occurrence

Most visible symptoms described in OSA adult population are daytime sleepiness and loud disruptive snoring. However, more of these symptoms might be associated with others type of sleep disorders or sleep-related breathing disorders, which may result in confounding factors for OSA detection. Furthermore, most of the patients are unaware of these symptoms since they require the identification by a bed partner or family member. This fact could explain the high percentage of undiagnosed cases in the population, where only 5% to 10% of the affected population ([Puertas et al., 2005](#)). In that context, the knowledge of OSA risk factors along with the suspect of some symptoms could help clinicians to detect and set priorities for OSA highest patients.

2.2.1 Craniofacial anatomy

The evolutionary changes in the human anatomical basis due to the acquisition of speech predisposed the upper airway to collapse during sleep. Anatomical changes such as the foreshortening of the maxilla, palate, ethmoid and mandible contributed to the shortening of the oral cavity and narrowing of the pharynx ([Davidson, 2003](#)). Besides these changes, there exist factors related to anatomical abnormalities in OSA subjects, being the most common the shortening and retrodisplacement of the maxilla and mandible ([Lowe et al., 1995](#); [Schwab et al., 2003](#)), and the displacement of the hyoid bone ([Guilleminault et al., 1984](#)). Consequently, the tongue, soft palate and soft tissue surrounding the UA are displaced posteriorly, which contribute to the narrowing of the oropharyngeal cavity. Soft tissues such as soft palate, tonsillar pillars, tongue, and uvula, along with an abnormal increase of their size due to inflammation or hypertrophy may also reduce the upper airway diameter enlargement [Ryan and Bradley \(2005\)](#). In the same manner, the upper airway (UA) structure configuration infers directly in OSA predisposition due to the high probability of pharynx's collapse. This effect is highly prevalent in snorers and OSA subjects, whose upper airway configuration causes an increased pharyngeal airflow resistance due to small pharyngeal airway lumen ([Bradley et al., 1986](#)).

2.2.2 Excess body weight

Among the important OSA risk factors, there are the central obesity (e.g. waist circumference, cervical perimeter) and excess of adipose tissue. Both conditions were reported in most of epidemiologic studies, being in many cases the 60% of patients referred for a diagnostic evaluation ([Strohl and Redline, 1996](#)). There are ways on how the excess of body weight can alter the normal

upper airway mechanics during sleep through several distinct mechanisms resulting in the upper airway obstruction: increased pharyngeal fat deposition, alterations in neural compensatory mechanisms, increased respiratory control system instability.

2.2.3 Ethnicity, age, and gender

Most of the studies carried on the OSA prevalence had been focused on Caucasian population (e.g. North America, Europe, or Australia). Nevertheless, several studies have started to focus on other ethnic groups such as Hispanic, African American and Asian. Considering the cephalometric differences in soft tissues among ethnic groups (Will et al., 1995) as well as the differences referred to body configuration, for instance Asians are less obese than Caucasians; the OSA risk is higher in a specific ethnic group rather than others, in particular, Asians have more pronounced skeletal abnormalities than Caucasians subjects, which are considered such as etiologic factors due to the increased risk and greater OSA severity in Asians despite lesser obesity (Lam et al., 2005).

Several studies have reported higher OSA prevalence in adult population over the age of 65 years rather than middle age (45-65 years) adult population, almost three times higher (Young et al., 2002a). Thus, it seems that aging has a positive correlation with the OSA predisposition. This fact is applied to middle age adult population (45-65 years), where the OSA prevalence progressively increases with age (Bixler et al., 1998). Moreover, the age-related increase in prevalence is caused by physiological changes such as fat deposition at the pharyngeal area and lengthening of the soft palate, which predispose the UA to blockage (Malhotra et al., 2006). We might expect aging leads to an increase in OSA evidence over the older age (65 years), however, this increase reaches a plateau after the age of 65 years (Young et al., 2002a). Despite the high OSA prevalence in the older adult population, it seems that there are other chronic diseases or sleep-disorders breathing that contribute to OSA predisposition, which suggests that OSA in older adults represents a distinct clinical entity than middle age adults.

The sex gender is another factor that contributes to OSA predisposition with a high prevalence in middle age men. Epidemiologic studies confirmed a high OSA prevalence in the adult male population, almost twice or three times prevalence than women. There are aspects that contribute to difference in prevalence, for instance classical symptoms such as loud snoring, nocturnal snorting or gasping are not reported in women (Young et al., 1996), whereas symptoms of fatigue and lack of energy they are usually reported (Chervin, 2000; Shepertycky et al., 2005). Also, the prevalence of many chronic disorders of middle and older age are higher in men than women (Waldron, 1985). In the same manner, the difference in upper airway shape, craniofacial

morphology, as well as the deposition of fat around the neck, could be considered as OSA risk factor in men. Beyond that, the number of undiagnosed cases is still high, in particular, in women ([Puertas et al., 2005](#)).

2.2.4 Occurrence

Most of the epidemiologic studies have been focused on characterizing the prevalence in different ethnic groups such as Caucasians, Asians, Hispanics, and African-Americans ([Punjabi, 2008](#)). As described in section [2.2.1](#), the anatomical-related differences among ethnic groups as well as the different severity symptoms such as snoring could contribute to the higher OSA prevalence in some ethnic groups rather than others ([Ong and Clerk, 1998](#)).

The common approach in OSA prevalence studies is referenced to gender and age, where the high prevalence is in adult male population: 2- to 3-fold greater risk for men as compared to women ([Strohl and Redline, 1996](#)). Also, the increase of OSA prevalence is correlated to aging in the middle age adults. In the case of elder age adults (65-90 years), the OSA prevalence is 3-fold higher than the prevalence estimates for OSA in middle age adults ([Lam et al., 2010](#)).

As regards women, it is known that classical symptoms such as loud snoring, nocturnal snorting or gasping are not usually reported in women ([Young et al., 1996](#)). Nevertheless, they seem to be uncertain since they might appear during the period of pregnancy ([Loube et al., 1996](#)). Hormonal function also plays a role in OSA pathogenesis, for instance the prevalence is higher in post-menopausal than pre-menopausal women ([Bixler et al., 2001](#)). To further complicate, the lack of awareness of the physicians due to a lower index of suspicion of OSA in women, it may contribute to the clinical under-recognition of OSA in women ([Dement and Leary, 2009](#)). Therefore, the screening of OSA-related symptoms such as snoring, fatigue, insomnia should be routinely conducted in women, considering the medical record of the patient ([Punjabi, 2008](#)).

2.3 Clinical diagnosis

Considering multiple risk factors play a role in OSA predisposition, the high undiagnosed cases, and comorbidities that patients could develop such as hypertension, it suggests that OSA diagnosis is a complex task. Therefore, it is necessary the development of adequate protocols to be able to identify the higher-risk patients and provide alternative treatment to lesser-risk patients. According to ([Puertas et al., 2005](#)), they propose a clinical exploration of the patient with clinical suspicion for OSA syndrome by means of an attending physician or pneumonologist. This exploration approaches the visual screening, clinical variables analysis, and questionnaires about

symptoms related to OSA. If the symptoms identified in the screening remain, then the patient will be referred to sleep disorders unit to proceed with the diagnosis.

The gold standard method for the OSA diagnosis is the polysomnography (PSG) test ([American Association of Respiratory Care-Association of Polysomnography, 1995](#)). It requires an overnight stay of the patient at the sleep disorders unit within a hospital with trained staff who are capable of monitoring and interpreting the real-time neurophysiological and cardiopulmonary variables in order to evaluate the quantity and quality of the sleep as well as the identification of different breathing events, heart rhythm, and limb movements. As a result of this test, the apnea-hypopnea index (AHI) is computed as the average number of apnea and hypopneas episodes per hour of sleep. According to the American Academy of Sleep Medicine (AASM), the classification of OSA severity considers both the AHI (i.e. overnight monitoring test) and the degree of daytime sleepiness ([Flemons et al., 1999](#)). The rating of severity for the syndrome is based on the most severe component. A detailed description of both is presented in Table 2.1.

Table 2.1: *Apnea-hypopnea index (AHI) thresholds and degree of daytime sleepiness for OSA severity assessment defined by American Academy of Sleep Medicine.*

Type	AHI threshold	Degree of daytime sleepiness
Mild	$5 \leq \text{AHI} \leq 15$	Unwanted sleepiness or involuntary sleep episodes during activities and require little attention (e.g. watching TV, reading).
Moderate	$15 < \text{AHI} \leq 30$	Unwanted sleepiness or involuntary sleep episodes during activities and require some attention (e.g. meetings, concerts).
Severe	$\text{AHI} > 30$	Unwanted sleepiness or involuntary sleep episodes during activities and require more active attention (e.g. driving, eating, walking).

Previous to diagnosis, there is a screening stage to give priority to higher-risk patients and refer them to sleep disorders units. Many screening methods have been developed such as those based on predictive models and questionnaires. In the case of predictive models, clinical features ([Hoffstein and Szalai, 1993](#)), craniofacial measurements ([Kushida et al., 1997](#)), and information collected by questionnaires (e.g. snoring, gasping at night, witnessed apneas) are usually used as predictors ([Maislin et al., 1995](#); [Viner et al., 1991](#)). As regards questionnaires, they consist of a series of questions used for OSA screening based on risk factors by means of a risk grouping system. Thus, the method predicts the likelihood of the patient having OSA. The most popular are: Berlin Questionnaires ([Netzer et al., 1999](#)), Epworth Sleepiness Scale ([Johns, 1991](#)), and the *STOP-BANG* questionnaire ([Chung et al., 2008](#)). In the same manner, there are also those based on visual inspection such as the Mallampati classification ([Mallampati et al.,](#)

1985). Although this method evaluate the ease of endotracheal intubation through the distance from the tongue base to the roof of the mouth, it is also used as screening method for OSA evaluation due to visual inspection of oropharyngeal cavity, where the more space, the lesser probability that UA collapses (Puertas et al., 2005). The Mallampati classification provides four scores (i.e. class I, II, III, IV) associated with the distance from the tongue base to the roof of the mouth. Thus, high score (class IV) means a high probability that UA collapses because of the narrowing of the oral cavity, consequently, a high predisposition for OSA. Besides those well-known screening methods, in the OSA literature, there is a variety of screenings devices such as oximeter (Wiltshire et al., 2001), electroencephalogram (Fischer et al., 2004), actigraphy recordings (Natale et al., 2012) and analysis of tracheal sound recordings (Reyes et al., 2014) by mobile Apps.

In summary, the variety of screening methods along with their low costs, fast screening process, and less invasive for the patient, make it suitable to be used for OSA screening, but the majority is still undergoing validation. Consequently, these screening methods could serve to aid the physicians in decision-making and should be used with prudence.

2.4 Physical and speech abnormalities in OSA

The acquisition of the speech on humans along with the language came with the species evolutions. This evolution was drawn on anatomical changes, specifically in the supralaryngeal vocal tract (SVT). These anatomical modifications include a shortening of the maxillary, ethmoid, palatal and mandibular bones; narrowing and elongation of the SVT; shortening and the lateral expansion of the face by rotation of the foramen magnum anteriorly, migration of the palate posteriorly, and shifting of the tongue into the oropharynx (Davidson, 2003).

The anatomical changes of the SVT provided humans with the ability to vocalize the vowels and consonants that constitute speech. The speech production begins at the larynx, then the SVT -above the larynx- acts as a tube, filtering the sound produced by the larynx. Then, the simultaneous change of positions of the pharynx, tongue, and lips will produce speech. However, the introduction of these anatomical modifications elicited an obstructing anatomy of STV, in particular, soft tissues obstruction (soft palate and nasal tissue) and the displacement of the tongue into the oropharynx. Thus, those modifications predispose the upper airway to collapse during sleep. Therefore, the evolution gave the ability to speak and develop language to humans, but predispose them to OSA development (Davidson, 2003).

2.4.1 Craniofacial and upper airway abnormalities

Besides the OSA evolutionary predisposition due to the speech acquisition, many studies have documented physical alterations in OSA subjects such as craniofacial abnormalities and enlargement of upper airway (UA) soft tissue structures using advanced imaging techniques.

In (Schwab et al., 2003), by means of sophisticated volumetric analysis techniques on magnetic resonance imaging (MRI), authors found that soft tissues structure surrounding the upper airway is enlarged in OSA subjects, in particular, the volume of the lateral pharyngeal walls (p -value < 0.0001), volume of the tongue (p -value < 0.0001), and total soft tissue (p -value < 0.0001). Lowe et al. (1995) reported differences between control and OSA subjects in craniofacial structures assessed by lateral cephalometry such as position of mandible (p -value < 0.05), mandibular plane angle (p -value < 0.05), total face height (p -value < 0.01), upper face height (p -value < 0.05), and lower face height (p -value < 0.05). In addition, they also reported differences in tongue volume (p -value < 0.001), soft palate size (p -value < 0.001), and upper airway size determined from computed tomography. Lyberg et al. (1989) studied oropharyngeal soft tissues using cephalometric analysis. They found significant differences between control and OSA subjects in length of the soft palate (p -value < 0.001), and distance of close contact between the tongue and the soft palate. Moreover, they found significant differences in the area of the soft palate, being greater in OSA subjects than control (p -value < 0.001). Guilleminault et al. (1984) obtained abnormal cephalometric measurements, in particular, the distance between the hyoid bone and the mandibular plane was longer in OSA subjects due to the position of the hyoid bone is frequently abnormally low in this group. This effect along with other craniomandibular abnormalities have been also reported in OSA diagnosed subjects in (Jamieson et al., 1986). By means of cephalometric landmarks analysis, they also found an abnormal lower position of the hyoid bone and more acute reposition of the mandible in OSA diagnosed subjects.

Despite the detailed examination of craniofacial and soft tissue structure surrounding the upper airway provided by advanced imaging techniques, these techniques are generally cost-expensive, time-consuming, radiation exposure, and invasive for patients. In this context, Lee et al. (2009a) proposed a novel method based on the anthropometry and photogrammetry resulting in faster and non-invasive method for craniofacial morphological assessment. Thus, the study compare the craniofacial phenotype of OSA ($AHI \geq 10$) and non-OSA ($AHI < 10$) subjects by contrasting 71 local features (i.e calibrated and uncalibrated measurements) extracted from facial landmarks previously annotated on frontal and profile digital photographs of the head.

As a result, the most of features showed significant differences between both contrasted groups. Among the most discriminative features, OSA subjects showed greater values in face width ($p\text{-value} < 0.001$), mandible width ($p\text{-value} < 0.001$), neck depth ($p\text{-value} < 0.001$), neck perimeter ($p\text{-value} < 0.001$), and cranial base area ($p\text{-value} < 0.001$). Besides the craniofacial phenotype characterization, [Lee et al. \(2010\)](#) proposed the hypothesis that the facial phenotype could be linked to the upper airway anatomy, in particular, surface facial assessment could provide extra information to craniofacial phenotype characterization. By means of the analysis of magnetic resonance imaging, the authors found correlations between craniofacial measurements and upper airway structures, for instance linear correlations: midface width and tongue volume ($r = 0.70$, $p\text{-value} < 0.001$), neck perimeter and soft palate volume ($r = 0.50$, $p\text{-value} < 0.001$), neck perimeter and tongue volume ($r = 0.76$, $p\text{-value} < 0.001$), and midface width and neck perimeter ($r = 0.81$, $p\text{-value} < 0.001$). Thus, these findings confirmed that the excess of adipose tissue surrounding the upper airway structure increases the OSA predisposition, consequently, it suggests that abnormal or particular speech patterns in OSA population may be expected from altered structure or function of the upper airway (UA).

2.4.2 Speech disorders

Early studies approached the study of speech anomalies in OSA subjects under the hypothesis the nature of dysfunction or anomaly should not just rely only on respiratory but also speech anomaly. According to ([Monoson and Fox, 1987](#)), the disturbances in abnormal speech are usually categorized as:

- **Articulation disorder:** it is characterized by a miss pronunciation of the speech sounds related to faulty functional or organic control of the tongue, lips, hard, and soft palate.
- **Resonance disorder:** it is characterized by an abnormality in vocal quality, typically related to coupling of the vocal tract (pharynx and oral cavity) to the nasal cavity. Resonance disorder can also be related to an abnormality in vocal tract damping and compliance.
- **Phonation disorder:** it is defined as abnormal sound production related to pharyngeal dysfunction.

[Monoson and Fox \(1987\)](#) carried out a preliminary perceptual study of speech disorders in 39 individuals: 13 obstructive-mixed sleep apnea, 13 chronic obstructive pulmonary disease (COPD), and 13 normal. The perceptual analysis task was carried out by three speech pathologists and consisted in listening subject's speech samples and categorize speech samples as normal

or abnormal regarding as articulation, phonation, and resonance. The subjects' speech samples were recordings of the first paragraph of the "rainbow passage", which contains a high percentage of occurrences of phonemes present in the Standard American English (Fairbanks, 1960). The speech samples were assigned a normal/abnormal rating based on the judgment of two out of the three speech pathologist. As results of listening task from 13 obstructive-mixed sleep apnea group 62% subjects were judged as abnormal, from 13 COPD group 23% subjects were judged as abnormal, whereas from 13 normal group 8% subjects were judged as abnormal. From obstructive-mixed sleep apnea subjects judged as abnormal, 87.5% of them were judged unanimously judged as abnormal, evidence of the larger percentage of inter-observer agreement for that group. The speech samples assigned as abnormal by speech pathologist were characterized as dysarthric with abnormal resonance. Since the etiology of dysarthria is associated with an injury to the central nervous system, and motor-speech disturbance of articulation, resonance, and phonation (Darley et al., 1969). This type of speech disorder is associated with slight "slurring" of the articulator or total immobility of tongue, velum, pharynx, and larynx. As regards abnormal resonance, the common etiology of resonance disorder is associated with airflow imbalance due to velopharyngeal dysfunction between the oral and nasal cavities, being perceived as hyponasality or hypernasality. However, speech pathologists did not describe the sleep apnea subjects as hypernasal or hyponasal. Some studies pointed out that the abnormal resonance in speech is associated with vocal tract damping characteristics instead of airflow imbalance between oral and nasal cavities. This type of resonance is called "cul-de-sac" (Peterson-Falzone, 1982). According to (Monoson and Fox, 1987), this type of resonance has been reported in structural and functional changes such as aberrant tongue posture, pharyngeal hypoplasia, and retrognathia, which have been also reported in apnea population.

Later, Fox et al. (1989) replicated this study but including more patients, 81 individuals: 27 obstructive-mixed sleep apnea, 27 chronic obstructive pulmonary disease (COPD), and 27 normal. Two approaches were carried out. First, to assess how the perceptual descriptors contributed to the perception of speech disorders by rating the degrees of abnormal, resonance and phonation found in subject's speech on a 1 to 7. After judgments of speech pathologists, the sleep apnea group was judged as abnormal for all three descriptors: articulation, phonation, and resonance. Second, by means of a discriminant function and using the speech disorder descriptors ratings as predictors, normal (healthy) and sleep apnea subjects were identified. The results of classification task were 63% of sleep apnea subjects correctly identified whereas 96% of normal subjects, even the most predictive variable among the descriptors was the phonation. Although its function is attributed to the larynx, which it has a limited role in the physiopathology of the

disease, an important finding was the roughness in the speech of apnea subjects, a perceptual characteristic related to lack of periodicity of the sound, thus disturbance in phonation patterns. It should be noted that this finding match to the first approach, where the phonation achieved a high degree in the sleep apnea group.

Both studies (Monoson and Fox, 1987; Fox et al., 1989) opened a new front for acoustic-analysis-based OSA detection (Fiz et al., 1993; Robb et al., 1997). In (Fiz et al., 1993), the harmonic structure of vowel spectra is analyzed extracted from sound recordings of five vowels (/a/, /e/, /i/, /o/, /u/) of 28 men (18 OSA subjects, 10 healthy subjects), finding a narrower frequency range and lower number of harmonics for vowel utterances /e/ and /i/ in OSA subjects. These findings suggest that there are differences in laryngeal behavior between OSA and non-OSA subjects. Since the population under study showed significant differences in body mass index (BMI), there was the hypothesis that the acoustic difference could be due to BMI differences between both groups rather than OSA. The excess of weight and the fat deposition around pharyngeal walls might influence the anatomical structure of the upper airway (UA), consequently, it might affect the speech production. However, split the population under study into two groups (normal weight and overweight), Fiz et al. (1993) did not find significant acoustic differences between two groups. In fact, it is important to point out that general obesity has a relevant impact on speech features, where its close relationship with OSA might contribute to the performance of speech features for OSA prediction. This aspect is also discussed thoroughly in Chapter 4.

In (Robb et al., 1997), since the OSA syndrome is a vocal tract phenomenon, the vocal tract behavior was studied by means formant frequencies and bandwidths analysis, which reflects the sound resonating properties of the vocal tract during speech production and measures the vocal tract compliance, respectively. Therefore, the vocal tract was evaluated in a group of 20 individuals (10 untreated diagnosed OSA, 10 non-OSA) by means of formants frequencies and bandwidth analysis extracted from vowels audio recordings: /a/, /i/, and /u/. Two hypotheses were addressed by Robb et al. (1997): first, in the basis of the findings in (Guilleminault et al., 1984), where they found an abnormally low position of the hyoid bone in OSA diagnosed subjects, this suggests that the vocal tract in OSA subjects are longer as compared to non-OSA subjects. Consequently, it is expected that OSA speakers produce lower formant frequencies, in particular, the first (F1) and second formant (F2). The results confirmed this hypothesis where the average values of F1 for vowels /i/ (p -value = 0.03) and /a/ (p -value = 0.05), and F2 for vowels /i/ (p -value = 0.05) and /u/ (p -value = 0.01) were significantly lower in OSA subjects as compared to the non-OSA subjects. Second, the increase of velar and pharyngeal compliance

have been reported in OSA diagnosed subjects (Brown et al., 1985), thus it is expected to alter vocal tract elasticity. As Robb et al. (1997) pointed out, excessive tissue compliance would tend to increase the sound damping within vocal tract. This increase is measured by formants bandwidths. An increase in sound damping will correspond to the larger bandwidths of the formants, consequently, wider formant bandwidths are expected in OSA subjects. The results showed significant differences in average formants bandwidths between the OSA and healthy subjects: /i/ (p -value = 0.05), /u/ (p -value = 0.05), /a/ (p -value = 0.01).

Based on early contributions, new approaches for automatic OSA detection were addressed by means of the use of automatic speech processing, in particular, the automatic speaker recognition techniques, and different machine learning techniques. These studies have been studied for Spanish (Fernández-Pozo et al., 2009) and Hebrew (Goldshtein et al., 2011) languages. Among the most used acoustic speaker features in these studies there are the cepstrums (Fernández-Pozo et al., 2009), formants frequency, and bandwidths (Montero-Benavides et al., 2016), and linear predictive coding (Goldshtein et al., 2011), all extracted by different types of speech (i.e., sustained and/or continues speech), along with different machine learning techniques such as classification models (Gaussian Mixture Model, Hidden-Markov Model, Support Vector Machine), and feature selection methods. Nevertheless, despite the good performance obtained by these studies, those results must be carefully regarded as there is the suspicion of machine-learning-related pitfalls which might explain the good performances obtained by those studies. This aspect is completely reviewed in Chapter 5.

2.5 Summary

In this chapter, a review of the Obstructive Sleep Apnea (OSA) syndrome has been presented. This syndrome is cataloged as a type of sleep-related breathing disorder, and it is produced by the blockage of the upper airway (UA) at the level of the pharynx due to the lack of muscular tone or fat deposition surrounding the pharyngeal cavity, resulting in breath cessation and a decrease in the level of the oxygen. The blockage of the UA could be total (apnea) or partial (hypopnea), in both with a breath's cessation episodes greater than 10 seconds. The aggravation of this condition relies on a group of factors which predispose the patient to suffer from this disease, being the main factors associated with anatomic abnormalities, age, and excess of weight, along with the medical comorbidities in OSA patients which could be the cause of mortality. Population-based epidemiologic studies pointed out a quite high prevalence in adult male as compared to women adult, in particular, in the middle age (45-65 years), where the

OSA prevalence shows a positive linear correlation with that age range.

The gold standard for OSA diagnosis is the polysomnography (PSG) test. As a result of this test, the apnea-hypopnea index (AHI) is computed as the average of apnea and hypopnea episodes per hour of sleep. Despite its high reliability, the costs introduced in the diagnosis are high related to lack of adequate equipment in the sleep disorders units, time-consuming, labor requirement, invasive for patients, and long waiting lists. As an alternative to this costly procedure, different methods of OSA screening have been explored such as questionnaires, clinical predictors, and screening portable devices in order to give priority to higher-risk patients and then refer them to PSG test, consequently, the diagnosis cost may reduce. Nevertheless, many studies agree that most of the screening methods are still no exhaustively validated, which might contribute to the increase of the undiagnosed cases and the diagnosis costs.

The final part of this chapter has been devoted to present a brief review of the connection between craniofacial abnormalities, speech disorders, and OSA along with a description of early studies carried out that demonstrate those connections.

Chapter 3

Databases and Methods

THIS CHAPTER describes the adopted experimental framework to assess and present the experiments and results contained in this Thesis. It includes the description of databases and supervised learning methods used for OSA prediction by means of the state-of-the-art speaker’s voice and craniofacial characterization.

First, we describe the speech recordings and photographic material of OSA patients used in the development of different training schemes, and the data capture protocol we followed. Second, the training methods are presented together with the performance evaluation and validation criteria.

3.1 Introduction

The exploration of speech and facial features for OSA prediction, in previous studies, has been applied over datasets that include audio recordings of sentences and sustained-vowels utterances, and medical images such as cephalometric radiograph, three-dimensional magnetic resonance imaging reconstruction of the upper airway (UA), as well as frontal-profile head digital photographs.

Early contributions to the state-of-the-art speech-based OSA detection are from the late 1980s ([Monoson and Fox, 1987](#); [Fox et al., 1989](#)), nevertheless, there were no references on the design and collection of databases including speech recordings from healthy and OSA patients in the literature until 2008 with the introduction of automatic voice characterization techniques for OSA assessing. During the period of 2007-2008, the first Spanish database was collected at Hospital Clínico de Málaga in Spain ([Fernández-Pozo et al., 2008](#)). Hereafter for the sake of completeness this database will be referred to as the Apnea Database v1.0. During the same

period, another Apnea-related database was collected in Israel (Zigel et al., 2008). Later, in 2010 the version v2.0 of the Apnea Database was released, and collected at Hospital Quirón Salud de Málaga. To date, the Apnea Database v2.0 data is still being collected. The collection of both Apnea databases was supervised by researchers from the Signal Processing Application group (GAPS) from the Universidad Politécnica de Madrid (UPM) and medical staff from the respective hospitals under the supervision of M.D. José Alcazar.

The most recent Spanish database reported was collected at Hospital Txagorritxu of Vitoria in 2012 (Durán-Cantolla, 2012). To our knowledge, no other databases focusing on the OSA syndrome have been identified in the literature. For the experiments of this Dissertation, we have used both versions of the Apnea Database (v1.0 and v2.0). The differences between both databases rely on the number of patients, the speech corpus, and the capture protocol, but the main differences rely on the subjects' health-condition and the inclusion of female subjects, which defines a remarkable difference between both datasets, as described below:

- **Apnea Database v1.0:** severe OSA diagnosed subjects ($AHI > 30$) belong to OSA group, whereas asymptomatic subjects along with healthy subjects and mild OSA patients ($AHI < 10$) belong to Control group.
- **Apnea Database v2.0:** female subjects were included and OSA ($AHI \geq 10$) and Control ($AHI < 10$) groups only include OSA-related symptomatic subjects.

As regards methods, this chapter also includes the description of supervised-learning models used for training models so as to assess the OSA severity in subjects by means of speech and craniofacial features. Despite the well-known performances on these methods, it must be carefully reviewed the training and testing procedures in order to avoid undesirable biases or over-fitting, which could lead to over-optimistic results. Thus, the methodology includes the description of validation methods used along with the methods for performance evaluation.

3.2 Databases

This section reviews the databases used for the experiments, in particular, both versions of the Apnea multimodal database. It should be noted that both databases were also described in previous Doctoral dissertations (Fernández-Pozo, 2011; Blanco-Murillo, 2013; Montero-Benavides, 2017) and research articles (Fernández-Pozo et al., 2009; Montero-Benavides et al., 2016; Espinoza-Cuadros et al., 2016). It should be noted that audio and images samples from new patients are still being collected in Apnea Database v2.0. Considering these points, we have reviewed the

main properties of both versions of Apnea Database, but we have mainly focused on the Apnea Database v2.0 since all experiments and analysis of this Dissertation were performed on it. In contrast, the Apnea Database v1.0 was used along with other databases as development dataset but not as training or testing dataset. This aspect is described in section 3.2.3.

3.2.1 Apnea Database v1.0

The Apnea Database v1.0 is a multimodal database and contains speech audio recordings and facial photographs of severe OSA subjects ($AHI > 30$), and healthy or mild OSA subjects ($AHI < 10$). These AHI thresholds used for OSA severity evaluation were specified in advance by the clinicians involved to focus on the particular AHI range in which they were interested. These thresholds were also used in the experiments in the Apnea Database v2.0.

The Apnea Database v1.0 was designed by the researchers from GAPS-UPM and the collection was carried out in the Respiratory Department at Hospital Clínico de Málaga (Spain). The design and collection procedure were first presented in (Fernández-Pozo et al., 2008), where the criteria for speech corpus and protocol capture were established based on the speech disorder descriptors reported by (Monoson and Fox, 1987) and (Fox et al., 1989). Most of the speakers are from Andalusia (southern Spain) and the database collection was carried out in two different sessions: the first session before the polysomnography test and the other after several months under CPAP treatment. The latter session was used by previous studies (Blanco-Murillo, 2013) to evaluate the progress of apnea voice characteristics for a particular patient before and after treatment.

3.2.1.1 Data collection procedure

The Apnea Database v1.0 was recorded in the Respiratory Department at Hospital Clínico de Málaga (Spain) from 2007 to 2008 under the supervision of an assistant and using the equipment provided by the team of researchers from GAPS-UPM.

The speakers were selected from the list of patients visiting the Pneumology Department of the Hospital Clínico de Málaga. All the speakers lived in the same area and exhibited a similar dialect, which makes this database homogeneous in terms of dialect and exhibited a similar dialect. The speakers were people who had experienced recurrent apnea episodes, often referred by their bed partners during a medical interview with the pneumonologist. By means of this interview clinical history and an exhaustive physical examination focusing on sleep-related symptoms, associated conditions, comorbidities and anthropometrics measures have been also collected. Subject's weight and height were recorded while wearing light clothes. Body mass

index (BMI) was calculated as the ratio of body weight (in kilograms) and the square of height (in meters). Cervical perimeter (in centimeters) was also measured at the level of cricothyroid membrane.

Based on this first screening, if the patient is likely to suffer from OSA, the patient is referred to the waiting list for polysomnography test before having a final diagnosis and an adequate treatment. Patients that were included in our database were selected according to the clinician's experience and additional criteria (e.g. similar age, body mass index, etc.) to ensure homogeneity among the two populations. The selected subjects were taken to a separate room with the adequate acoustic condition and the recording equipment for collecting speech previously patient's consent.

After the speech material had been recorded, the patient was referred to the overnight polysomnography test for the final diagnosis. If the final diagnosis gave an Apnea-Hypopnea Index (AHI) over 30, the patient was assigned to the OSA group, otherwise to the control group. Patients with severe OSA were treated by CPAP. After three months of treatment, the patients were asked to perform a second recording session with the purpose of assessing any alterations the CPAP treatment could introduce. According to [Blanco-Murillo \(2013\)](#), there were no differences between both speech recording sessions, being the intra-speaker variability due to CPAP treatment negligible. It is important to point out that second recording session was not used in this Thesis.

3.2.1.2 Speech Corpus

The speech corpus consists of four sentences in Spanish repeated three times and a sustained vowel /a/ also repeated three times. Early acoustic-analysis-based studies for OSA detection were based on sustained vowels recordings ([Fiz et al., 1993](#); [Robb et al., 1997](#)). The analysis performed on these linguistic units could be useful since the speech production variability is minimized and is easier to track the speech production behavior as compared to continuous speech in OSA diagnosed subjects, but the information related to speech production as articulation is dismissed, which could contain more information of speech production related to OSA acoustic features. Therefore, the speech corpus also includes four sentences, which were designed trying to cover relevant linguistic/phonetic context related to peculiarities in OSA voices derived from the acoustic ([Fox et al., 1989](#)). These peculiarities include abnormalities in resonance, articulation, and phonation. Consequently, the speech corpus design includes instances of the following specific phonetic and linguistic contexts:

- **Phonation anomalies:** sentences include continuous voice sounds to explore irregular phonation patterns such as lack of periodicity that could be related to fatigue and respiratory troubles in apnea patients (Payne et al., 2006).
- **Articulatory anomalies:** according to (Fox et al., 1989), the speech of OSA subjects was described as dysarthric with abnormal resonance. The dysarthria is a type of speech disorder and it is characterized by the lack of muscular control over the speech mechanism and the grade of severity of this disorder could lead to slurred speech due to the total immobility of the velum, tongue, pharynx, and larynx, thus, an abnormal articulation patterns are expected in subjects with this type of speech disorder. The sentences also include continuous voice sounds affected by certain preceding phonemes that have their primary locus articulation near the back of the oral cavity such as the guttural sounds as Spanish voiced and voiceless velar plosives /g/ and /k/, respectively.
- **Resonance anomalies:** in (Monoson and Fox, 1987) the speech pathologists described abnormal resonance patterns in obstructive sleep apnea subjects speech due to airflow imbalance as consequence of velopharyngeal dysfunction between the oral and nasal cavities being perceived as hyponasality or hypernasality. The speech pathologists perceived a different type of resonance, coined as “cul-de-sac” resonance, which might be related to increases in pharyngeal compliance in OSA subjects along with an aberrant posture of the tongue, pharyngeal hypoplasia, and crowding of the pharynx. Considering hyponasal and hypernasal characteristics, the sentences were designed that allow intra-speaker variation measurements such as the degree of vowel nasalization measurement in different vowel contexts. As regards vocal tract damping characteristics, the sentences include specific allophones of the vowel phonemes where may occur some “cul-de-sac” resonances in OSA subjects (Robb et al., 1997).

To make this Thesis self-contained, a review of the design and criteria of four sentences are presented below, in particular, relevant aspects of the Thesis development. A detailed description of these sentences can be found in (Fernández-Pozo, 2011; Blanco-Murillo, 2013). In Figure 3.1, the four sentences are presented.

Sentence 1	Francia, Suiza y Hungría, ya hicieron causa común
IPA transcription	fraNθja sujθa i ungría dZa jθjeronj kawsa komun
Translation	France, Switzerland and Hungary, already made common cause
Sentence 2	Julián no vio la manga roja, que ellos buscan, en ningún almacén
IPA transcription	xuljan no βjo la maŋma roxa ke eloz βuskan en niŋgun almaθen
Translation	Julián did not see the red sleeve, that they seek, in any store
Sentence 3	Juan no puso la taza rota, que tanto le gusta, en el aljibe
IPA transcription	xwan no puso la taθa ʔota ke taNto leχusta en el alxiβe
Translation	Juan did not put the broken cup, he is so fond of, in the tank
Sentence 4	Miguel y Manu llamarán entre ocho y nueve y media
IPA transcription	miyel i manu λamaran eNtre otʃo i nweβe i meðja
Translation	Miguel and Manu will call between eight and half past nine

Figure 3.1: Sentences of Apnea Database, including IPA phonetic transcription and English translation.

A brief description of sentences in Figure 3.1 is detailed below:

- **Sentence 1:** it was taken from ALBAYZIN database (Moreno-Bilbao et al., 1993). This sentence contains a sequence of successive /a/ and /i/ vowel sounds in the words “Francia” [ʼf r a N θ j a] and “Suiza” [ʼs u j θ a]. According to (Robb et al., 1997), differences in first formant (F1) and second formant (F2) of vowels /a/ and /i/ were found between OSA and non-OSA subjects; these differences are shown as abnormal resonance in the former group. In addition, it also contains guttural sound as voiced (/g/, “Hungría” [u η ʼg r i a]) and voiceless velar plosives (/k/, “causa” [k a w s a] and “común” [k o m u n]), where abnormal articulation may be expected in OSA subjects (Fox et al., 1989).
- **Sentence 2 and 3:** on the basis of intra-speaker differences between nasal and non-nasal vowels due to OSA-related velopharyngeal dysfunction found by Monoson and Fox (1987), the analysis of intra-speaker variation vowel of nasalization in nasal and non-nasal context are approached by both sentences. Some examples of these contrastive pairs is found in nasal context as /a/ of “manga” [ʼm a η g a] and neutral context /a/ of “taza” [ʼt a θ a].
- **Sentence 4:** it includes a relatively long sentence mainly composed of voiced sounds read as one single melodic group. Since the sleep apnea speakers usually present fatigue in upper airway muscles due to breathing disorders, this sentence may be helpful to discover phonation anomalies during the sustained generation of voice sounds, or the presence of differences in speech dynamics, in particular, pause patterns, which may be related

to abnormal articulation patterns (e.g. dysarthria). In addition, the sentence includes several sounds embedded in a nasal context that could be used to study phonation and the articulation in nasalized vowels.

The four sentences length is 4 seconds on average and has 42 phones, although they exhibit a correct grammatical structure they lack full meaning. This fact may affect the data collection since the sentences uttered in front of a microphone are almost always read in a fairly unnatural way due to stress since the task is strange to most people and requiring additional cognitive processing which could mislead the speaker. In contrast, the sustained vowels could not be affected by these drawbacks and could be useful to study phonation measurements that detect the variation in amplitude and frequency of the speakers, but in a more controlled way.

Despite the sentences were designed to capture information related to speech disorders in OSA population, the speaking style and the dialect could also affect this search. For instance, in the Andalusian dialect, the nasalization of vowels could lead to false evidence of abnormal speech patterns as well as the loss of pronunciation of intervocalic consonant such as /n/, and aspiration of implosives consonants such as /g/ ([Alvar-Ezquerro, 1959](#)) that could overlap the abnormalities in resonance.

3.2.1.3 Speech and image data recording

Speech samples were recorded in an acoustically isolated room with a sampling frequency of 50 kHz and encoded in 16 bits. Afterward, it was downsampled to 16 kHz before processing.

The recording equipment consisted of a standard laptop computer equipped with a *500 Plantronic* headset microphone with A/D conversion and digital data exchange through a USB-Port. A control acquisition process software installed on the laptop, developed at Universidad Politécnica de Madrid (UPM), was used to guide the speaker and help the assistant in controlling the process. The assistant was requested to control the software, place the microphone, and check that the signals were not saturated. The assistant was asked to request as many repetitions as necessary to avoid misreading and any other spurious phenomena.

The patients were requested to read the sentences listed in Figure 3.1 in a comfortable and natural mode in three satisfactory repetitions of each of the four sentences, observing the specific rhythmic structure previously indicated and under the assistant's supervision. In the same manner, the patients were asked to emit the sustained vowel /a/ for at least one second. This procedure was repeated four times. In addition, the recordings were carried out in two different sessions. In the first session, the sleep apnea group was recorded before being diagnosed, at the

same time the control group (i.e. asymptomatic subjects along with non-OSA subjects) was also recorded. The second session, the sleep apnea group was recorded after several months under CPAP treatment. In our experiments, only the first recording session was included in the development dataset, as described in section 3.2.3.

As regards image data recording, frontal and profile digital photographs of the head were obtained for each subject in the database under controlled illumination conditions and over a white flat background. A conventional digital data was used to obtain images with a size of 2272×1704 and a color depth of 24 bits.

It is important to point out that this photographic dataset was not used in our experiments for OSA prediction since photographs belong to subjects which were included in the development dataset. This aspect is described in detail in section 3.2.3.

3.2.2 Apnea Database v2.0

After the first evaluations on the Apnea Database v1.0, many questions arose regarding the informativeness of the sustained vowels and the possibility to extend the analysis to other groups such as female patients, young and middle age adults, more than one dialect, and alternative grouping of speakers according to AHI. Consequently, the design of the Apnea Database v2.0 aimed to improve some of the weakness related to the first version.

The data collection started in 2010 at the Hospital Quirón Salud de Málaga (Spain) under the supervision of M.D. José-Alcazar. Currently, the data recording of new patients is still in progress, as well as new features and changes have been incorporated such as audio recordings of phrases and vowels in two different positions (sitting and supine), breathing audio recordings in wakefulness, and audio recording of interview between patient and the assistant.

3.2.2.1 Data collection procedure

The data recording was carried out at the Sleep Disorders Unit of Hospital Quirón Salud at Málaga (Spain). Although we followed the same data collection procedure of Apnea database v1.0, there are remarkable differences, in particular, those concerning to subjects' clinical diagnosis, as described in section 3.1, as well as different exclusion criteria.

In addition, subjects in the dataset referred typical OSA symptoms during the preliminary interview with a pneumonologist, such as excessive daytime sleepiness, snoring, choking during sleep, or somnolent driving. It is important to point out that subjects underwent a detailed interview concerning their clinical history and a thorough physical examination focusing on sleep-related symptoms, associated conditions, comorbidities, and anthropometric measures.

As regards exclusion criteria, weight and height homogeneity criterion were not included in this database. Thus, the extent of the analysis to other groups related to age, BMI is possible. In addition, information related to symptoms and clinical condition as snoring, smoking, hypertension, loss of concentration and upper airway condition were collected as part of patients' clinical history.

Since the data collection is still in progress, it could be interesting to briefly review the history of this database along with the development of our research carried out on this topic. At the beginning of 2014, data from 650 subjects under evaluation and suspected to suffer from sleep-disordered breathing was collected and included in the Apnea Database v2.0. First studies and experiments on this database were conducted, whose results were published in (Espinoza-Cuadros et al., 2015; Montero-Benavides et al., 2016). Later, at the end of 2015, additional data from 166 subjects were included in the Apnea Database v2.0. Thus, the previous experimental framework was replicated into the updated database along with the development of new experiments, whose results were recently published in (Espinoza-Cuadros et al., 2016; Tyan et al., 2017).

Currently, the Apnea Database v2.0 contains audio recordings and photographs of 816 subjects (233 women and 583 men). As compared to v1.0, the v2.0 includes the whole range of AHI's values, ranging from 0 to 108. Similar to the v1.0, the AHI's values were arranged according to the interest expressed by the clinicians. Accordingly, three OSA severity classes were defined: healthy and mild OSA ($AHI < 10$), moderate OSA patient ($10 \leq AHI \leq 30$), and severe OSA ($AHI > 30$). As regards inclusion and exclusion criteria, since the audio recordings include more than one dialect, subjects with no Andalusian dialect were excluded. In the same manner, subjects with photograph capture errors and unavailable clinical variables were also excluded. Consequently, 729 subjects were validated (204 women and 525 men). It should be noted that those excluded patients were included in development dataset, as described in section 3.2.3. Descriptive statistics on the female and male population are presented in Tables 3.1 and 3.2.

3.2.2.2 Speech and photographic data recording

Besides the four sentences, the Apnea Database v2.0 includes the speech recordings of five Spanish sustained vowels: /a/, /e/, /i/, /o/, and /u/. Since the spectral analysis of speaker's voice over sustained vowels are less affected than a continuous speech by sources of variability, for instance intonation and differences in dialect, the sustained vowels makes suitable for this type of analysis such as formant frequencies and bandwidths analysis (Robb et al., 1997; Montero-

Table 3.1: Summary on the male subjects, including the mean and standard deviation ($\pm SD$) for each of the clinical variables collected along with speech and image records. Control ($AHI < 10$), mild OSA ($10 \leq AHI \leq 30$) and severe OSA ($AHI > 30$) are considered jointly and separately for each variable. The last column includes p -values estimated by non-parametric Wilcoxon rank-sum test applied to each variable considering only the control and OSA groups.

Clinical variables	All (n=525)	Control (n=160)	Mild OSA (n=216)	Severe OSA (n=149)	p -value
AHI	22.63 $\pm$ 18.43	4.86 $\pm$ 2.94	19.22 $\pm$ 6.30	46.64 $\pm$ 14.07	< 0.01
Age (years)	49.01 $\pm$ 12.34	44.22 $\pm$ 11.47	51.55 $\pm$ 12.48	50.48 $\pm$ 11.64	< 0.01
Weight (kg)	90.95 $\pm$ 16.95	84.26 $\pm$ 12.68	90.44 $\pm$ 14.89	98.87 $\pm$ 20.25	< 0.01
Height (cm)	175.17 $\pm$ 7.16	175.33 $\pm$ 7.70	174.92 $\pm$ 6.86	175.38 $\pm$ 7.03	NS
BMI (kg/m ²)	29.60 $\pm$ 5.02	27.41 $\pm$ 3.73	29.54 $\pm$ 4.44	32.06 $\pm$ 5.86	< 0.01
CP (cm)	42.00 $\pm$ 3.17	40.55 $\pm$ 2.79	42.06 $\pm$ 2.89	43.47 $\pm$ 3.28	< 0.01

Abbreviation: BMI, Body Mass Index; CP, cervical perimeter; NS, nonsignificant.

Table 3.2: Summary on the female subjects, including the mean and standard deviation ($\pm SD$) for each of the clinical variables collected along with speech and image records. Control ($AHI < 10$), mild OSA ($10 \leq AHI \leq 30$) and severe OSA ($AHI > 30$) are considered jointly and separately for each variable. The last column includes the p -values estimated by non-parametric Wilcoxon rank-sum test applied to each variable considering only the control and OSA groups.

Clinical variables	All (n=204)	Control (n=108)	Mild OSA (n=65)	Severe OSA (n=31)	p -value
AHI	15.11 $\pm$ 17.52	4.38 $\pm$ 3.05	17.60 $\pm$ 6.26	47.29 $\pm$ 21.50	< 0.01
Age (years)	51.42 $\pm$ 11.65	47.69 $\pm$ 11.46	55.78 $\pm$ 10.69	55.25 $\pm$ 10.00	< 0.01
Weight (kg)	78.72 $\pm$ 17.50	76.38 $\pm$ 17.53	80.07 $\pm$ 17.21	84.03 $\pm$ 17.09	0.04
Height (cm)	161.16 $\pm$ 6.36	161.84 $\pm$ 6.49	160.61 $\pm$ 5.76	159.96 $\pm$ 6.99	NS
BMI (kg/m ²)	30.37 $\pm$ 6.86	29.18 $\pm$ 6.59	31.15 $\pm$ 7.06	32.90 $\pm$ 6.62	< 0.01
CP (cm)	36.85 $\pm$ 2.93	36.14 $\pm$ 2.56	37.54 $\pm$ 2.86	37.87 $\pm$ 3.66	< 0.01

Abbreviation: BMI, Body Mass Index; CP, cervical perimeter; NS, nonsignificant.

[Benavides et al., 2016](#)).

The speech was recorded at a sampling frequency of 50 kHz and encoded in 16 bits. Afterward, it was downsampled to 16 kHz before processing. Recordings were made in a room with low noise and patients at an upright or seated position. At the end of 2014, audio recordings in supine positions were incorporated. Recording equipment was a standard laptop computer equipped with a *500 Plantronics* headset microphone that includes integrated A/D conversion and digital data exchange through USB-port. At the beginning of 2015, the interview between the clinical assistant and patients, and breathing recordings were also collected.

Besides subjects' clinical diagnosis, another remarkable difference as compared to v1.0 is speakers were asked to read their phrases and vowels in two different positions: sitting and lying on their back (supine position). This decision is based on the analysis of differences related to UA configuration, which would vary from one position to the other. Since the apnea episodes occur during sleep, the objective is reproducing the same conditions during recordings in order to detect OSA characteristics patterns ([Durán-Cantolla, 2012](#)).

Therefore, in this Thesis, the four sentences have been used in the experiments. It is important to point out that since 2015 to date only the phrases 2 and 4 are collected. In addition, the speech recordings of subjects on supine position started at the beginning of 2015. Currently,

there are not enough samples of speech in supine position, consequently, they were not employed in the experiments of this Thesis. A detailed description of the number of utterances of each type of sentence in Apnea Database v2.0 is presented in Table 3.3.

Table 3.3: *Number of utterances for each type of sentence in Apnea Database v2.0.*

Gender	Sentences				Total
	1	2	3	4	
Female	182	204	182	204	772
Male	455	525	455	525	1960

In addition, since each speaker pronounces the designed set of sentences, thus the number of speech samples used for the experiments does not match the number of speakers. In Tables 3.4 and 3.5, the number of utterances for control and OSA groups is presented in detail for both genders.

Table 3.4: *Number of utterances for Control and OSA groups in male population in terms of each type of sentence.*

	All	Control	Mild OSA	Severe OSA
Sentence				
1	455	136	193	126
2	525	160	216	149
3	455	136	193	126
4	525	160	216	149
Total	1960	592		1368

Comparing Tables 3.1 and 3.2, the number of samples in OSA (mild and severe) and Control groups are not balanced in men, but in the female population, they are equally balanced. The number of subjects with mild or severe OSA ($AHI \geq 10$) is approximately twice than the Control group ($AHI < 10$). This information plays an important role in the training of regression and classification models in order to predict the OSA severity in subjects.

As regards photographic data, frontal and profile digital photographs of the head were ob-

Table 3.5: *Number of utterances for Control and OSA groups in female population in terms of each type of sentence.*

	All	Control	Mild OSA	Severe OSA
Sentence				
1	182	97	57	28
2	204	108	65	31
3	182	97	57	28
4	204	108	65	31
Total	772	410		362

tained before the speech recordings at the same hospital room over a white flat background. As compared to v1.0, no special actions were taken beyond a simple control for patients' front and profile photographs and some instructions to guarantee that the neck area is visible in the profile image. No calibration action for allowing the conversion from pixel measurements to metric dimensions (e.g. measuring the distance from the camera) was considered, and the manual identification (e.g. by palpation) of facial landmarks was also avoided. A standard *Logitech QuickCam Pro 5000* webcam was used to collect images with a size of 640×480 pixels and a color depth of 24 bits.

Figure 3.2: *Frontal and profile photographic samples from the Apnea Database v2.0.*

3.2.3 Additional databases

The representation of the acoustic information of a speaker utterance recording by means of the state-of-the-art speaker recognition techniques is usually represented as a sequence of feature vectors corresponding to the short-term spectral envelope of the embedded sounds (see Chapter 4). Due to the natural variability in speech production, different utterances will exhibit sequences of vectors with different lengths.

Classical speaker modeling can be performed by two different forms: nonparametric such as Vector quantization (Soong et al., 1987) and Dynamic Time Warping (Furui, 1981), and

parametric models such as Gaussian Mixture Models ([Reynolds et al., 2000](#)), and Hidden Markov Models ([Naik et al., 1989](#)). In the case of the parametric models, the sequence of feature vectors is generally represented into fixed-length vectors representing all the relevant acoustic information in the utterance. This representation is performed by training an acoustic subspace model from the utterance’s sequence of vectors and extracting the model’s statistic from the subspace model. This acoustic subspace model represents the total acoustic variability observed in a large population, that is the primary directions of total acoustic variability. A description in detail of the whole speech-based OSA recognition is presented in Chapter 4.

Therefore, the development dataset consists of different Spanish spoken databases, covering a wide range of phonetic variability from continuous/read Spanish speech. Speech samples were recorded in an acoustically isolated room at a sampling of 50kHz and encoded in 16 bits by using a headset microphone. Afterward, they were downsampled to 16 kHz before processing. Most of the speech samples of all databases were recorded under controlled noise conditions. Moreover, the Apnea Database v1.0 and discarded samples of Apnea Database v2.0 were also included in the development dataset. It is important to point that those samples were not used for training and testing. Since most of these databases contain both-gender speech recordings, thus, we considered a gender-dependent approach since we aim to evaluate, separately, the speech features in OSA female and male subjects. A general description of databases, included in the development dataset, is detailed below along with the number of samples for each database is shown in Table 3.6.

- **ALBAYZIN database** ([Moreno-Bilbao et al., 1993](#)) is a Spanish spoken database for speech recognition purposes. The database is divided into three parts: a phonetic database, an application database, and a Lombard speech database. For this Thesis, we used the phonetic and application database. The corpus for the phonetic database consists of a set of phonetically balanced sentences whereas the application database consists of sentences related to retrieval information from a geographical database. In total, there are 14525 Spanish utterances of 301 speakers (female and male) from 18 to 55 years old and an average length utterance of 3 seconds.
- **AT&T Bell Labs Headlines database** ([Alvarez-Cercadillo et al., 1995](#)) is a Spanish headline speech database for continuous speech recognition research. The corpus consists of 650 sentences that have been pronounced by 51 speakers (25 female and 26 male). The total number of utterances are 10439.
- **NUMLIM Telefonica I+D database** ([López-González et al., 1998](#)) is a Spanish spoken

database and its corpus consists of connected numbers. The number of utterances is 5270 that have been pronounced by 95 speakers (32 females and 63 male).

- **VESLIM Telefonica I+D database** (Toledano et al., 2003) is a Spanish spoken database and its corpus consists of 10 multiword utterances and 75 isolated word utterances. The speeches are from 10 different dialects from Spain. This corpus is phonetically balanced. For this Thesis, we used the 10 multiword utterances for each of 1012 speakers (415 female and 597 male).
- **TRESVEL Telefonica I+D database** (Martinez et al., 1997) is a Spanish spoken database designed to study, evaluate and compensate the effect of speech rate on Large Vocabulary Continuous Speech Recognition (LVCSR) systems. Its corpus consists of slow, average and fast speech. For this Thesis, we used the average speech and it is composed 3200 utterances pronounced by 40 speakers (20 female and 20 male).
- **APNEA database v2.0**: some patients were included in development dataset since they have not all the clinical parameters or some photograph capture errors. They are 15 women and 33 men and the total number of utterances are 401. Adding this apnea dataset into development dataset contributes to the total acoustic variability subspace. Accordingly, it includes part of the acoustic variability of subjects with different OSA severity levels. It is important to point out that these patients have not been involved in the training and testing process.
- **APNEA database v1.0**: all samples from this database were not used for the experiments due to symptoms-related variation between Control and OSA groups.

Table 3.6: *Description of databases included in the development dataset. It includes the number of speakers, the number of utterances pronounced by these speakers, and the number of hours of recordings*

Database	Female			Male		
	# spk.	# utt.	hrs.	# spk.	# utt.	hrs.
ALBAYZIN	152	6800	6.51	149	7725	7.18
BELL LABS	25	4998	6.32	26	5441	6.67
NUMLIM	32	1778	1.79	63	3492	3.37
VESLIM	415	4150	8.99	597	5970	12.55
TRESVEL	20	1600	1.67	20	1600	1.66
APNEA 1.0	-	-	-	82	984	1.41
APNEA 2.0	15	96	0.15	33	305	0.49
Total	659	19422	25.43	970	25517	33.33

3.3 Methods

In this section, we describe the two approaches performed to evaluate the speech and craniofacial features for predicting OSA severity in patients. In the same manner, we describe the supervised learning models used along with the description of training, testing, and validating schemes, and the metrics used in performance evaluation.

Besides the speech craniofacial features extraction techniques used, which is described in Chapter 4, the modeling of the speech feature requires the use of statistical modeling techniques, in particular, the well-known Gaussian Mixture Models extensively used in the state-of-the-art speech and speaker recognition, hence, it is also briefly described in this section.

3.3.1 Problem formulation

According to the objectives of this Thesis, the main goal aims to explore the speech characterization techniques employed in state-of-the-art speaker recognition for OSA severity prediction in subjects. In the same manner, we also aim for craniofacial phenotype characterization in OSA subjects through the evaluation of different facial characterization techniques employed in the state-of-the-art face recognition. It is important to remark that our objective is not the development of an automatic OSA recognition system, as carried out by previous studies, but the thorough evaluation of proposed features, in particular, speech features, whose complexity makes them difficult to assess by statistical methods. Consequently, in order to explore both types of features for OSA assessment, we adopted two different approaches: regression for AHI prediction and classification for OSA severity classification. In addition, it is known the good performances reported by speech features for speaker's physical characteristics such as age (Bahari et al., 2014), height (Poorjam et al., 2015b), and weight (Poorjam et al., 2015a). Therefore, we have also evaluated the speech features for predicting the clinical variables to validate the feature extraction procedure.

3.3.1.1 Regression approach

The AHI prediction can be formulated as a machine learning regression problem as follows: given a training dataset S_{tr}^j of speech recordings, facial photographs and Apnea-Hypopnea Index (AHI) along with the clinical variables information:

$$S_{tr}^j = \{x_n, y_n^j\}_{n=1}^N \quad (3.1)$$

where $x_n \in \mathbb{R}^p$ denotes the feature vector representation for nth utterance or facial image of

the training dataset and $y_n^j \in \mathbb{R}$ denotes the corresponding value of the clinical variable for the speaker of that feature vector representation; j corresponds to AHI or a particular variable in the set of V clinical variables ($j = 1, 2, \dots, V$; i.e age, height, weight, BMI, and cervical perimeter).

The goal is to design an estimator function f^j for each clinical variable, such that, from a set of speech/facial feature vectors from an unseen testing subject x_{test} , the difference between the estimated value of the particular clinical variable $\hat{y}^j = f^j(x_{test})$ and its true value of y^j is minimized. In the case of speech features, the predicted variable $\hat{y}^j$ is the mean of all prediction related to each utterance feature vector as depicted in Figure 3.3.

Figure 3.3: Block diagram describing Apnea-Hypopnea Index (AHI) and clinical variables estimation for a test subject by means of the regression approach and using speech features.

Besides of predicting the AHI from speech/facial samples, we have also tested the performance when using the same techniques to estimate the other clinical variables (age, height, weight, BMI and cervical perimeter). We think this evaluation is relevant for two main reasons: first, to validate our methodology, by comparing our results estimating age, height, and weight with those previously reported over general speaker populations such as age (Bahari et al., 2014), height (Poorjam et al., 2015b), and weight (Poorjam et al., 2015a). Second, to identify correlations between speech and other clinical variables that can increase the likelihood of false discovery based on spurious or indirect associations between these clinical variables and AHI. This second aspect is described in chapter 5, where we critically review previous approaches to OSA assessment.

3.3.1.2 Classification approach

Despite we might know how close or far our prediction method is from the gold standard diagnosis method, nevertheless, from a medical approach, only the predicted AHI value might not seem to give enough information about how well do the features and the prediction model system discriminate among different OSA severity conditions as compared to the gold standard method.

A straightforward approach for this issue might be the OSA severity classification according to the predicted AHI value by setting a threshold value. However, this address could lead to misclassification results or false discoveries as there are some machine-learning aspects that need

to be addressed carefully such as differences in the number of samples between OSA severity groups. Consequently, the best option is to perform the OSA severity classification by means of classification models properly trained instead of the straightforward approach.

Given a training dataset of speech and facial recordings $S_{tr} = \{x_n, y_n\}_{n=1}^N$ and OSA severity class information $y_n \in \{1, -1\}$; being a binary classification: -1 for low OSA ($AHI < 10$) and 1 for mild-severe OSA ($AHI \geq 10$).

The goal is to train a classification model, such that, from a set of speech/facial feature vectors from an unseen testing subject x_{test} , the detection error associated with the predicted class could be the minimum along with high generalization model performance.

Since the dataset is imbalanced, this drawback is outperformed by adding regularization terms to model cost function in order to penalize the majority class. This aspect is described in detail in section 3.3.2.

In the case of speech features, the final decision is performed over the mean of all utterance scores as depicted in Figure 3.4. It is important to point out that the final class of each utterance provided by the classification model was not considered as the evaluation was performed on a subject instead of a single utterance. The search for an optimal decision threshold is described in section 3.3.4.2.

Figure 3.4: Block diagram describing OSA severity assessment by means of a classification model and using speech features.

3.3.2 Supervised learning models

3.3.2.1 Support Vector Machine

Support Vector Machine (SVM) (Cortes and Vapnik, 1995) is a supervised learning model based on minimum risk optimization, which aims for establishing a high-dimensional optimal separation boundary between two classes. Due to its flexibility and their well-known performance in a variety of problems, they have been used in the last years, for instance in automatic speaker recognition context by means of spectra features (Campbell et al., 2006b) and in automatic face detection (Osuna et al., 1997).

The SVM approach is based on mapping the non-linearly separable input feature space onto a high-dimensional feature space where it can be linearly separable by means of a hyperplane (see Figure 3.5). Consequently, the SVM performs the classification task by determining the maximum margin separation hyperplane between two classes, which is $\frac{2}{\|w\|}$. This hyperplane can be expressed in the primal form of SVM by means of a discriminant function:

$$f(x) = w^T \varphi(x) + b \quad (3.2)$$

where $\varphi(x)$ is a mapping function in the input feature space, w is a row vector with the same dimension of $\varphi(x)$, and b is the bias term. The finding of the optimal hyperplane relies in the maximization of the margin $\frac{2}{\|w\|}$, therefore minimizing $\|w\|$. This is rewritten as the primal optimization problem as presented in the following expression:

$$\begin{aligned} \underset{w, \xi}{\text{minimize}} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \quad (3.3)$$

where $\frac{1}{2} w^T w$ is related to the maximization of the margin hyperplane (ξ_i), called slack variables, and determines the distance of training samples to the respective classes boundary. An error occur when the ξ_i exceed the unity, as depicted in Figure 3.5. $C > 0$ is the penalty parameter of the error term, also called regularization term. Therefore, the generalization performance of the SVM depends on a good setting of these parameters, in particular, the parameter C , which affects the trade-off between the margin (complexity) and the size of the slack variables ξ_i (i.e. a proportion of non-separable samples).

Once the mapping onto a high-dimensional space is done, we could proceed to solve the optimization problem in the new feature space $\varphi(x)$. However, it could become computationally infeasible due to the high dimensionality in the new feature space. Thus, the explicitly performing operation in high-dimensional space is avoided by means of a function called *kernel function* K , which is designed to meet the *Mercer condition* (Burges, 1998), and it can be expressed as an inner product of a mapping function φ in the form:

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle \quad (3.4)$$

The SVM learns in the high dimensional feature space given φ , but without ever having to explicitly find or represent vectors $\varphi(x)$. This shortcut is commonly known as the *kernel trick*.

Figure 3.5: Representation of the SVM underlying idea (left) and representation of a separable data classification and basic elements of the SVM (right).

Considering this shortcut, the solution of the optimization problem in the equation 3.3, can be solved more efficiently by introducing a dual set of variables and solving a dual optimization problem. Therefore, the discriminant function 3.2 can be rewritten as:

$$f(x) = \sum_{i=1}^{N_{sv}} \alpha_i t_i K(x, x_i) + b \quad (3.5)$$

where t_i are the truth output values $+1$ or -1 . x_i are the support vectors associated with maximum margin hyperplane, N_{sv} is the number of support vectors, α_i their corresponding weights and b the bias term estimated from the optimization process. Finally, the decision function is performed by the *sign* on the discriminant function (3.5), as described in expression 3.6.

$$\text{sgn}(f(x)) = \begin{cases} -1 & \text{if } f(x) < 0 \\ 0 & \text{if } f(x) = 0 \\ 1 & \text{if } f(x) > 0 \end{cases} \quad (3.6)$$

Besides these parameters, the kernel parameters also play a role in the generalization performance of the SVM. According to decision function 3.6, the SVM uses as a threshold a value equal to zero. It is important to consider this aspect since the optimal performance of a classifier relies on the decision of the threshold point.

A typical issue in classification approach is to deal with imbalanced data. Although there are methods in the literature that overcome this problem such as resampling and ensemble learning,

their performance might be limited due to the size of the dataset. As an alternative, several supervised learning models as SVM can deal efficiently in an internal way. In the case of SVM, it is sensitive to class imbalance due to the soft margin objective function (equation 3.3) applies the same penalty parameter (C) for both classes. This would cause the separating hyperplane to be skewed towards the minority class, which would finally elicit a suboptimal model. To overcome this problem, Veropoulos et al. (1999) proposed a modification of the SVM soft margin objective function by adding two regularization terms, such that C^+ is the penalization for the positive class examples, whereas C^- is the penalization for the negative class. Thus, the equation in 3.3 is rewritten as,

$$\begin{aligned} \underset{w, \xi}{\text{minimize}} \quad & \frac{1}{2} w^T w + C^+ \sum_{i=1}^N \xi_i + C^- \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{3.7}$$

For instance, if we assume positive class to be the minority class and negative class to be the majority class, thus by assigning a higher misclassification cost for the minority class examples than the majority class examples (i.e. $C^+ > C^-$), the effect of class imbalance could be minimized. As a rule of thumb, Akbani et al. (2004) reported that the reasonably good classification results could be obtained by setting the $\frac{C^-}{C^+}$ equals to the minority to majority class ratio. In the Thesis context, as the number of samples in OSA male group (majority class, C^+) is twice than the Control group (minority class, C^-), the regularizations terms can be expressed as shown in 3.7. As regards female population, according to Table 3.2, both classes are balanced.

$$\frac{C^+}{C^-} = \frac{1}{2} \tag{3.8}$$

3.3.2.2 Support Vector Regression

Introduced by Vapnik (2013) and also called ϵ -SVR, the Support Vector Regression (SVR) is a function approximation approach developed as a regression version of the Support Vector Machine (SVM) classifier. Similar to the SVM, the input space first is mapped onto a high-dimensional space by means of a nonlinear mapping function φ . Once the mapping is done, the goal is to find the optimal regression hyperplane $f(x)$ in which most of the training samples lie within an ϵ -margin (ϵ -insensitive zone) around this hyperplane, as depicted in Figure 3.6 (right). The regression hyperplane $f(x)$ linear can be expressed as the discriminant function equation

3.5.

Figure 3.6: Representation of the SVR underlying idea (left) and the soft margin loss setting for a linear SVR (right). Adapted from (Schölkopf and Smola, 2002).

Similar to the Support Vector Machine (SVM), the optimal regression hyperplane finding depends on the solution of the optimization problem as the equation in 3.3 and using the loss function called ϵ -insensitive loss function (Vapnik, 2013) described by

$$L(y, f(x)) = \begin{cases} 0 & |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon & \text{otherwise} \end{cases} \quad (3.9)$$

And the optimization problem can be expressed as:

$$\begin{aligned} & \underset{w, \xi}{\text{minimize}} && \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i + \xi_i^* \\ & \text{subject to} && y_i - w^T \phi(x_i) - b \geq \epsilon + \xi_i \\ & && w^T \phi(x_i) + b - y_i \geq \epsilon + \xi_i^* \\ & && \xi_i, \xi_i^* \geq 0 \end{aligned} \quad (3.10)$$

where ϵ controls the width of the ϵ -insensitive zone, which determines the level of accuracy of an approximation function and relies entirely on the target values of the training dataset. The parameter $C > 0$, as SVM, determines the trade-off between the model complexity and the degree to which deviations larger than the ϵ -insensitive zone are tolerated in the optimization of the hyperplane. This corresponds to dealing with a ϵ -insensitive loss function, where the deviations are penalized in a linear fashion (see Figure 3.6). Similar to SVM, the kernel parameter also

plays a role in the generalization performance of the SVR.

3.3.3 Model training and validation scheme

The training of a classifier or a predictive model consists in several steps: a) choosing a method analysis, b) feature selection, c) training the classifier or predictive model, d) validating the classifier, e) model performance evaluation (Cyran et al., 2013). Each step may introduce bias and errors in the developing process. According to (Foster et al., 2014), these aspects were not correctly managed in most of the studies published related to the use of machine learning techniques to develop classifiers for detection or diagnosis disease, obtaining a limited diagnosis and methods that are prone to over-fitting.

As described in section 3.3.1, the problem formulation aims to evaluate the speech and facial features for OSA assessment. Therefore, the problem formulation is turned into a machine-learning problem, where the correct address of common machine-learning pitfalls is required in order to avoid false discoveries. These common issues and relevant aspects are described below:

1. **Small number of samples:** most of previous speech-based OSA detection studies have been carried out over a small number of samples. Working in a high-dimensional feature space with a small number of training samples could lead to over-fitting, losing performance generalization of the model. For instance, at the beginning of the Thesis research, the maximum number of patients reported in those previous studies was 248 adult patients (Solé-Casals et al., 2014).
2. **High dimensionality data:** this aspect becomes critical if the problem includes small datasets, that is a small number of samples, as described in 1), there is a high chance to fall in over-fitting. Many of the previous studies for speech-based OSA detection did not correctly address this issue, where prediction/classification models are trained on a high dimensionality feature space and achieving high performances for predicting OSA.
3. **Feature selection:** considering a scenario with points 1) and 2), applying feature selection will lead to false discoveries due to the poor generalization of the trained model. Despite some studies conducted validation methods to avoid overfitting, the validation process performed in those studies is not clearly detailed, suggesting the appearance of false discoveries. For instance, the training, testing and validation process are performed on small datasets.
4. **Imbalanced dataset:** it is a common issue into machine learning applications such as

medical diagnosis and biomedical engineering due to the imbalanced classes in the dataset, for instance differences in prevalence between disease and control class (Foster et al., 2014). To overcome this inconvenient, there are many approaches that can address this problem such as cross-validation schemes, ensemble algorithms, and cost functions to penalize the majority class.

5. **Model hyperparameters tuning:** this aspect is also critical for classification or predictive model performances since it plays a role in the trade-off between generalization and complexity of the model. Therefore, the searching for optimal hyperparameters must be performed following a correct validation scheme so as to achieve the best trade-off.

3.3.3.1 K-Fold Cross Validation and Grid search

In this Dissertation, we have designed a scheme for developing and validating classification and regression models based on the k-fold cross-validation. In the same manner, we have followed a similar approach to (Hsu et al., 2003), called “Grid Search”, in order to search for the optimal hyperparameters of the models. Both processes are depicted in Figure 3.7 and described below:

- **K-fold cross validation:** to guarantee that all speakers were involved in the test, the dataset was split into k equal sized subsamples with no speakers in common. Then, of the k subsamples, a single subsample is retained for testing and the remaining $k - 1$ subsamples are used as training dataset and for “Grid Search”.
- **Grid search:** in order to find the optimum complexity of the model, we use the remaining $k - 1$ subsamples for searching for the optimal hyperparameters of the models. For this purpose, we implement a “Grid Search”, and consist in an exhaustive search through a specified set of hyperparameters using 5-fold cross validation. Therefore, various subsets of values are tested and the one with best results is picked. As regard classification approach, the best results correspond to the highest Area Under the Curve (AUC) value obtained. As regards regression approach, the best results correspond to the lowest Mean Absolute Error (MAE) obtained. Those performance metrics are described in section 3.3.4.1. A description of the Grid Search Process is presented in Table 3.7.

3.3.4 Evaluation of performance

In this section, the performance measures used for evaluating the proposed scheme for OSA assessment by means of two approaches are described.

Figure 3.7: Block diagram describing k-fold cross-validation and grid-search process for SVM/SVR model.

In the context of the medical diagnosis, the advances in medicine it usually leads to the development of new methods with many advantages over the reference one (gold standard method). Nevertheless, the evaluation of the test method should not rely on the prediction of the disease (Bland and Altman, 1986), but the degree of agreement between the reference and the proposed method. In the OSA context, we aim to measure the degree of agreement between the PSG and our OSA prediction model.

Moreover, the analysis of continuous variables (e.g voice or facial features, clinical variables, AHI, etc) was carried out. Therefore, a description of statistical analysis used in the experiments is provided.

3.3.4.1 Evaluation in Regression approach

Given the predicted AHI value or other clinical variables (age, height, BMI, weight and cervical perimeter) we measure both the Mean Absolute Error (MAE) and the Pearson Correlation (r).

Table 3.7: *Description of Grid search for selecting the optimal hyperparameters for a classification model. The same process is applied for regression, where the Mean Absolute Error (MAE) is used instead of the Area Under the Curve (AUC).*

```

k := 5 %number of folds
Otrain := {O1, O2, ..., Ok} %training observations
C := {2-20, 2-19, ..., 220} %values of model parameters
AUCglobal := {}

for each parameter i Ci in parameter set C:
  for each fold j in k
    modelj=TrainingModel(k-1 subsamples) %training model with k - 1 subsamples
    AUCj=modelj(1 subsample) %testing model with 1 subsample
  end
  AUCglobal=max(AUCj)
end

[AUCmax, index] = max(AUCglobal)
return C(index)

```

The Mean Absolute Error (MAE) provides the average absolute difference between actual y_i and estimated $\hat{y}_i$ values, and Pearson Correlation (r) evaluates their linear relationship between both values. As correlation coefficients can be very small values, we considered informative to report p -values for correlation coefficients as the probability that they were in fact zero (null-hypothesis). In addition, the results were also compared with the most basic estimator that is the average value of the training data, consequently, this estimator will reach a poor performance. Hereafter it is called prior.

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (3.11)$$

In the case of regression from speech samples, since each patient pronounces four or two utterances (see Table 3.4), different AHI or clinical values will be estimated for the same subject. Therefore, as the final predicted value for each subject, the mean of all predicted values is given, as described above,

$$\hat{y}_{subject}^j = \frac{1}{N_{utt}} \sum_{i=1}^{N_{utt}} \hat{y}_i^j \quad (3.12)$$

where N_{utt} is the number of utterances per subject, and j corresponds to a particular variable (j =AHI, age, BMI, weight, cervical perimeter, height).

3.3.4.2 Evaluation in Classification approach

Given speech/images samples Y of an unseen subject, the model will provide a decision on whether it is likely or not that the subject actually suffers from OSA. In a mathematical form, the problem can be defined such as a likelihood ratio (Λ) between the likelihoods obtained for the null-hypothesis (H_0) to be that the patient does not suffer from OSA and the alternative hypothesis (H_1) to be that patient does, as shown in equation 3.13,

$$\Lambda = \frac{P(H_1 | Y)}{P(H_0 | Y)} \leq \lambda_{TH} \quad (3.13)$$

where λ_{TH} is a threshold for determining the OSA severity from the scores (λ) obtained from the likelihood ratio (Λ). The equation can be rewritten as:

$$\Lambda(Y) = \frac{P(Y | H_1)P(H_1)}{P(Y | H_0)P(H_0)} \leq \lambda_{TH} \quad (3.14)$$

where the knowledge of the priors $P(H_1)$ and $P(H_0)$ is essential for the solution in equation 3.14. These priors are the probability that we get a subject who actually suffers from OSA relies directly on the prevalence that we are considering as a reference. For instance, if we take the prevalence of the syndrome in male population, then $P(H_1)$ should be approximately 2% to 4%, whereas if we take as a reference only the patients that seem to be suffering from some kind of this disorder and decide to visit the doctor, the priors might be different. In this Thesis, we have assumed that the priors of both events are equal, consequently, the equation 3.14 can be simplified, avoiding prior probabilities as:

$$\Lambda(X) = \frac{P(X | H_1)}{P(X | H_0)} \leq \lambda_{TH} \quad (3.15)$$

As shown in equation 3.15, according to the likelihood ratio value (Λ) of a test sample (X), and the decision threshold (λ_{TH}) can be classified as either that patient does not suffer from OSA (T_0) or patient suffers from OSA (T_1). The overall performance of test samples will be a set of likelihood ratios called scores.

Therefore, from the N test samples scores ($\lambda_i, i = 1, 2, \dots, N$), obtained from the likelihood ratio (Λ) evaluated in each test sample (X_i), and following a non-discrete sense, two probability distributions are expected: $P(\lambda_i | H_1)$, related to OSA group and $P(\lambda_i | H_0)$ for Control group. These probability functions are depicted in Figure 3.8 through their density function: $f(\lambda_i | H_1)$, related to OSA group and $f(\lambda_i | H_0)$ for Control group. As depicted in Figure 3.8, the selected threshold (λ_{TH}) defines two types of errors (FP, FN), and they are described below:

Figure 3.8: Example of probability density function of scores resulting from binary classification. H_0 is the null-hypothesis, to be that the subject does not suffer from OSA whereas H_1 is the alternative-hypothesis, to be that the subject does. The value of the threshold λ will define the False Positives (FP) and False Negatives (FN).

- **False Positive (FP):** also called **Type-I error**, occurs when the null-hypothesis (H_0) is true, but it is rejected in favor of the alternative hypothesis (H_1). In other words, this type of error is produced when a test sample, who actually does not suffer from OSA, the system classifies it as OSA sample (T_1). In terms of rate (i.e. False Positive Rate, FPR), this type of error is defined through its density function as,

$$FPR = \int_{-\infty}^{\lambda_{TH}} f(\lambda | H_1) d\lambda \quad (3.16)$$

- **False Negative (FN):** also called **Type-II error**, occurs when the alternative hypothesis (H_1) is true, but it is rejected in favor of the null-hypothesis (H_0). In other words, this type of error is produced when a test sample, who actually suffers from OSA, the system erroneously fails in classifying as Control sample (T_0). In terms of rate (i.e. False Positive Rate, FPR), this type of error is defined through its density function as,

$$FNR = \int_{\lambda_{TH}}^{\infty} f(\lambda | H_0) d\lambda \quad (3.17)$$

As we can see in equations 3.16 and 3.17, these types of errors rely on the selected threshold (λ_{TH}). In the diagnostic of OSA context, we desire to minimize the False Negative Rate (Type-II error) to minimize number of undiagnosed cases that actually suffer from OSA. Consequently,

besides increasing the costs of treatment of this disease, the risk of having an undiagnosed subject will be detrimental to the health condition of the subject. In the same manner, the Type-I error also affects the system performance, but we assume that from a medical approach having false diagnosed patients (false positives cases) will not harm the health condition status of that patients whereas the Type-II error does. In other words, while performing the classification task, the test aims to minimize the probability of a OSA patient (H_1) not be diagnosed as a healthy subject (T_0) whereas the medical expert should warranty that the chances of healthy people (H_0) being wrongly diagnosed (T_1) are minimized.

As depicted in Figure 3.8, besides the False Negatives (FN) and False Positives (FP) samples error, there are also the True Negatives samples (TN), when negative test sample (H_0) is predicted as negative sample (T_0), and the True Positives samples (TP), when positive test sample (H_1) is predicted as positive sample (T_1). The classification performance can be measured by these four values through a confusion matrix, as shown in Table 3.8. A description of classification performance measurements used in the experimental part is described below:

- **Sensitivity:** the probability that a test result is positive when the actual condition is present. $Sens = \frac{TP}{TP+FN}$
- **Specificity:** the probability that a test result is negative when the actual condition is not present. $Spec = \frac{TN}{TN+FP}$
- **Accuracy:** the probability that a test outcome matches with the actual references results. $Acc = \frac{TP+TN}{TP+FN+FP+TN}$

Table 3.8: *Confusion matrix for OSA detection test*

		Polisomnography	
		OSA	Control
Screening test	OSA	TP	FP
	Control	FN	TN

Nevertheless, none of these indicators by itself represent the test's discriminatory performance. For instance, a high sensitivity may be accompanied by low specificity. As a complementary measure to these indicators, an additional performance metric was used, which relies on the Sensitivity and the Specificity. This measure is the **Diagnostic Odds Ratio (DOR)**. This type of measure is frequently reported in biomedical studies as an indicator of diagnostic performance (Glas et al., 2003) and it is defined as the ratio of the odds of positivity in disease relative to the odds of positivity in the no diseased. According to (Bland and Altman, 2000),

the odds is the ratio of the probability that the event occurs to the probability that it does not, in other words, it is the ratio of the number of times that the event of interest occurs to the number of times that it does not, as described in expression 3.18.

$$DOR = \frac{odds(\text{detected positive cases in diseased group})}{odds(\text{detected positive cases in non-diseased group})} \quad (3.18)$$

For a better understanding, considering the confusion matrix in Table 3.8 and OSA detection scenario, on the one hand, the odds of detected OSA cases within the actual OSA cases is the ratio of the number of times that OSA cases were detected in OSA group (TP) to the number of times that it does not (FN). On the other hand, the odds of detected OSA cases within the actual Control cases is the ratio of the number of times that OSA cases were detected in Control group (FP), to the number of times that it does not (TN).

In a medical diagnosis context, the priority is to minimize the probability of type-error II in case of a high-risk disease. Therefore, the higher the DOR, the better discriminative model.

$$DOR = \frac{TP/FN}{FP/TN} = \frac{TP \times TN}{FN \times FP} \quad (3.19)$$

In addition, the values of DOR ranges from 0 to infinity, with higher values indicating better discriminatory test performance. A value of 1 means that the test does not discriminate between patients with the disorder and those without it. Moreover, the DOR is usually reported with confidence intervals. Since the DOR ranges from 0 to infinity, it presents a skew distribution, thus confidence interval is estimated over the natural log of DOR. A 95% confidence interval for the log DOR is obtained as 1.96 standard error on either side of the estimate. The standard error is estimated by the following expression:

$$SE(\log DOR) = \sqrt{\frac{1}{TP} + \frac{1}{FP} + \frac{1}{FN} + \frac{1}{TN}} \quad (3.20)$$

It is important to point out that these classification performance measures metrics are threshold-dependent. For instance, for the SVM the performance measures rely on a threshold value equal to zero since its decision function is based on the *sign* function, as described in section 3.3.2.1. Besides that, a wider view of the classifier performance can be also achieved. This is done by describing the system performance for all possible threshold points. One of the most used parametric representation is the **Receiver Operating Curve (ROC)** (see Figure 3.9), which represent the model performance as a trade-off between the two types of error (Type-I error, Type-II error). As a single measure to compare different models' performances the **Area**

Under the Curve (AUC) is given. The AUC is the integral behind the ROC curve, and its values range from 0 to 1. An AUC of 1 indicates a perfect performance, whereas an AUC of 0 indicates a classifier with a completely wrong outcome. Moreover, an AUC of below 0.5 indicates random guessing of the classifier, which means a poor performance. A complete review of ROC graphs can be found in (Fawcett, 2004).

Figure 3.9: Description of the optimal values of sensibility, specificity, and threshold estimation in the ROC curve by searching for the maximum value within filtered sensibility values greater than 70%.

In order to generate a ROC we need to generate scores from a classifier rather than just a class label. There are many ways of producing such scores through the instance statistics of the classifier keep. For instance, in the Random Forest method the scores could be obtained by the proportion of votes of most voted class. In the Thesis context, the scores were obtained from the decision function of the SVM (see equation 3.5) and normalized (i.e. $[0, 1]$) afterwards.

Due to the nature of the SVM model, it provides neither score nor probability that a sample belongs to a class, it only provides the predicted class but defined in a specific threshold, which is 0 according to the decision function in 3.6. As a solution to this issue, the normalized values (i.e. $[0, 1]$) provided by the discriminant function of the SVM model (see expression 3.5) were used as scores in order to generate the ROC.

In summary, given a fixed threshold the classification performance can be reported at that point by different measures (e.g. accuracy, sensibility, specificity, DOR), as well as, all possible operating points represented by a single value; the Area Under the Curve (AUC). In a medical context, a single measure might be desired in order to get a straightforward interpretation of the results. This single measure could be the optimal operating point from the ROC curve. The choice of the optimal threshold relies on the minimization of the type of expected error, in

the case of the OSA diagnosis, considering the risk factors and the comorbidities, the goal aims to maximize the *sensitivity* at the expense of *specificity* in order to get lower False Negative rate (Type-II error), and increase the detection of true positive samples. The optimal threshold might be chosen so that the classifier gives the best trade-off between the costs of failing to detect positives samples against the cost of failing negative samples. It is common to assume equals costs for both, but in a real scenario, this is not always true. As described previously, in this Thesis, the misclassifying costs are not considered since they should be chosen by a clinician according to the relevance of each these errors.

As an alternative, a straightforward method (Figure 3.9) is proposed in this Thesis and consists of the following steps:

1. Searching for the optimal threshold through the ROC where the *sensitivity* and *specificity* reach together optimal rates by setting only a minimum rate of *specificity* of 70%. Therefore, *sensitivity* rates greater or equal than 70% are filtered along with their corresponding specificity rates.
2. From filtered *sensitivity* and *specificity* values, we choose those with minimum difference between them.

By means of these conditions, we guarantee the minimization of Type-II error through the maximization of the *sensitivity*, as well as both *sensitivity* and *specificity* reach balanced rates as we are dealing with an imbalanced dataset.

Figure 3.10: Bland-Altman plot for average PEFR by two methods. Extracted from (Bland and Altman, 1986).

3.3.4.3 Agreement between two methods

In medical diagnosis, it is frequently to compare two different methods of indirectly measuring some quantity such as blood pressure, gestational age, or cardiac stroke volume since the direct measurement is difficult or impossible in some cases. Therefore, an alternative method has to be evaluated by comparing measurements obtained by the new method and the gold standard technique. If the new method agrees sufficiently well with the established technique, the old method may be replaced, indeed whether this new method is cheap, quick and produces answers that agree with those from an established method sufficiently well for clinical purposes.

The correlation might be a measurement to assess the agreement between both methods, however, a question may arise: does correlation already assess the agreement between both variables or methods? Although some studies suggest the use of correlation coefficients as an indicator of an agreement, it is not always a good choice. According to (Bland and Altman, 1986), correlations depend on the range of the true samples. For instance, if the range is wide the correlation will be greater than if it is narrow. Only a perfect agreement between two methods will be associated with high correlation if only the points lie along the line of equality.

As an alternative, Bland and Altman (1986) proposed a visual approach in order to assess the agreement between two methods, and consist in plotting the difference between the methods against their mean as they are depicted in Figure 3.12, where the limits of agreement is defined as $\bar{d} \pm 2s$, $\bar{d}$ is the mean difference and s the standard deviation of the differences. Most of the differences (i.e. 95% of cases) lies between these limits of agreement. The agreement between the two methods can be clearly visualized by this method instead of the correlation through a scatter plot, as described in (Bland and Altman, 1986). In this Thesis, we want to assess the degree of agreement between the predictive model and the gold standard diagnostic test (PSG). Plotting the data by means of the Bland-Altman method, we obtain a view of agreement between the two methods and the values range that could lie near the mean and within the limits of agreement.

In the OSA diagnosis context, it might be ideal to obtain a measurement of the agreement between the PSG and the predictive model instead of only a visual approach. This fact was considered in (White et al., 1995), where the authors assessed the accuracy of a novel portable device to monitor sleep and breathing in the home. To determine the frequency of diagnostic errors encountered by the predictive model, assuming that the gold standard method is the PSG, by setting an AHI threshold, thus sensitivity and specificity could be obtained. However, the use of sensitivity and specificity as validation tool of a new diagnostic approach could be flawed

due to two reasons: 1) there is no well-defined “cut-off” above which sleep apnea is diagnosed and below which it is excluded. 2) small differences in AHI between systems, for instance, one or two events per hour, which are clinically unimportant, may be interpreted as total diagnostic errors by means of specificity and sensitivity evaluations.

As a result, [White et al. \(1995\)](#) developed a new approach for determination of diagnostic errors. In [Table 3.9](#), a description of this approach is given. It should be noted that we have included a new diagnostic approach called “no agreement”, not included in ([White et al., 1995](#)).

Table 3.9: *Description of diagnostic agreement approach for determination of diagnostic errors between PSG and test methods.*

Diagnostic approach	Condition
Agreement	$ AHI_{PSG} - AHI_{test} \leq 10 \ \& \ AHI_{PSG} < 40$ or $AHI_{test} \geq 40 \ \& \ AHI_{PSG} \geq 40$
Overestimate	$AHI_{test} - AHI_{PSG} > 10 \ \& \ AHI_{PSG} < 40$
Underestimate	$AHI_{PSG} - AHI_{test} > 10 \ \& \ AHI_{PSG} < 40$
No agreement	None of the previous conditions are accomplished

3.3.4.4 Statistical analysis

To perform the statistical analysis of continuous variables (e.g voice or facial features, clinical variables, AHI, etc) Student’s t-tests were performed for comparing normally distributed continuous variables. All data are expressed as mean $\pm$ standard deviation (SD). In addition, analyses of linear relationships among continuous variables were also performed using Pearson correlation. A p -value < 0.05 was considered statistically significant. All p -values below 0.10 are shown in the results, and those that are greater than or equal to 0.10 are shown as non-significant (NS), with the exception those values that provide extra information to the required task. In addition, the strength of the correlation coefficient is described as in ([Fowler et al., 1998](#)), that is values between 0 and 0.19 are referred to as very weak, between 0.2 and 0.39 as weak, 0.40 and 0.69 as modest, 0.70 and 0.89 as strong correlation and in the range 0.90 and 1.00 as very strong.

3.3.5 Gaussian Mixture Models in Speaker Recognition

The Gaussian Mixture Models (GMMs) framework have been widely used in the state-of-the-art text-independent speaker recognition with short-term spectral features ([Reynolds et al., 2000](#)) for many years until the development of factor analysis techniques ([Kenny et al., 2008](#)). This scheme can be seen as a likelihood ratio detector between a GMM target model and a speaker-independent GMM model, also called Universal Background Model (UBM). This GMM model is

trained with speech samples belonging to different speakers to represent as much as possible the speaker-independent distribution of the feature vectors. It is used as a prior to obtain specific target GMM models via Maximum a Posteriori Adaptation (see section 3.3.5.2). To obtain a similarity measure between test feature vectors and a given target model, a likelihood ratio is established between the likelihoods obtained for the GMM target and the UBM model.

3.3.5.1 Definition

A GMM (λ) is a stochastic model composed by a weighted sum of K finites mixture of D-multivariate Gaussian densities as given by the equation,

$$p(o_t|\lambda) = \sum_{i=1}^K w_i g(o_t|\mu_i, \Sigma_i) \quad (3.21)$$

where o_t is a D-dimensional vector, w_i the mixture weights, μ_i the GMM means, Σ_i the GMM variance and $g(o_t|\mu_i, \Sigma_i)$ is a D-variate normal distribution, with probability density function of the form,

$$g(o_t|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{(-\frac{1}{2}(o_t-\mu_i)^T \Sigma_i^{-1} (o_t-\mu_i))} \quad (3.22)$$

For example, Figure 3.11 shows a 4-components GMM trained from 2-dimensional feature (a) vectors and its contours in 2D (b).

Figure 3.11: A GMM of 4-components of 2-dimensional Gaussians (a) and their contours (b).

3.3.5.2 GMM training and MAP adaptation

Training a GMM model consists of estimating the parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}_1^K$ from a set of training observations. This is usually done by means of the Maximum Likelihood estimation, through the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Previous to the EM, a clustering stage via K-means is normally performed to favor a quick convergence of the EM algorithm (Linde et al., 1980).

In the state-of-the-art speaker recognition, in general, the number of speech samples available per speaker is not enough to properly train a GMM. As a solution to this, Reynolds et al. (2000) proposed the approach of a GMM-UBM framework. The underlying idea is to derive the speaker's model by updating the UBMs parameters (w_i, μ_i, Σ_i) via an algorithm called Maximum A Posteriori (Gauvain and Lee, 1994). This process is represented in Figure 3.6. Although all parameters of the GMM-UBM could be adapted, in the case of the state-of-the-art speaker recognition no significant outperforms in the global performance by updating all the parameters were reported. Moreover, the adaptation of GMM covariance matrices Σ_i could become computationally costly, slowing the adaptation process. Consequently, the GMMs means μ_i is only adapted.

The MAP estimation is a two-step estimation process. First, by Expectation-Maximization the sufficient statistic (see equations 3.24, 3.25, 3.26) of the training data are computed for each component in the UBM model. Second, as described in the expression 3.23, the “new” sufficient statistic estimates are then combined with the “old” sufficient statistic.

The specific of the adaptation process are as follows. Given a training dataset of observations, $O = O_1, O_2, \dots, O_N$, represented by a GMM ($\lambda = \{w_i, \mu_i, \Sigma_i\}_1^K$) and a UBM (λ_{UBM}) both of K components, $\{c_1, c_2, c_3, \dots, c_K\}$, the adapted mean vectors can be seen as a trade-off between the UBM model means and GMM (μ) in the form,

$$\mu'_i = \alpha_i f_i + (1 - \alpha_i) \mu_i \quad (3.23)$$

where

$$\alpha_i = \frac{n_i}{n_i + \tau} \quad (3.24)$$

$$n_i = \frac{1}{n_i} \sum_{t=1}^N P(C_i | o_t, \lambda_{UBM}) \quad (3.25)$$

$$f_i = \frac{1}{n_i} \sum_{t=1}^N P(C_i | o_t, \lambda_{UBM}) o_t \quad (3.26)$$

Being, C_i the i component of the UBM, α_i the adaptation coefficient, τ the relevance MAP factor, which controls the importance of training samples and the UBM within the adaptation procedure, and n_i and f_i the so-called 0^{th} and 1^{st} order statistic, respectively.

Figure 3.12: A Maximum Posteriori Adaptation process representation where a speaker model is adapted from a UBM. Adapted from (Reynolds et al., 2000).

In order to obtain a similarity measure between test feature vectors and a given GMM target, a likelihood ratio is established between the likelihoods obtained for the GMM target and the UBM. Therefore, the final score produced from a test observations set $O = O_1, O_2, \dots, O_T$ is computed as a likelihood ratio between the target model λ and the UBM λ_{UBM} . Taking logs this takes the form,

$$L(O, \lambda, \lambda_{UBM}) = \frac{1}{T} \sum_{t=1}^T (\log p(o_t | \lambda) - \log(p_t | \lambda_{UBM})) \quad (3.27)$$

3.4 Summary

In this chapter, the experimental framework has been extensively described, which includes the description of the databases and the methods used for OSA assessment by speech and facial features.

Both Apnea Databases (v1.0 and v2.0) were presented, including a review of the phonetic content of its corpus. Speech records and facial photography from subjects derived from pneumonologist, which show OSA-related symptoms, were collected just before going into polysomnog-

raphy test along with some additional information on patient's age, height, weight, and cervical perimeter. Additionally, other speech databases are described and they were used as development dataset during the experiments, as described in Chapter 4.

Furthermore, the formulation of the problem considering both approaches regression and classification was described. This formulation consists of the description of how the detection scheme is approach by following an automatic detection approach. Although the main goal does not pursue the design of an automatic OSA recognition, we followed this approach in order to assess the discriminative capability of the speech and facial features in OSA population. Moreover, the supervised learning methods used in the experimental part were also described, in particular, the aforementioned Support Vector Machine (SVM) and Support Vector Regression (SVR). Due to the size of the dataset, we have proposed a scheme for developing and validating the trained model.

Finally, a brief description of GMMs in the state-of-the-art speaker recognition was provided to get a better understanding of speech feature extraction process used along with the limits of utterances representation due to the natural variability in the speech production.

Chapter 4

Speech characterization in Obstructive Sleep Apnea

ONCE THE LITERATURE IN THE OBSTRUCTIVE SLEEP APNEA has been described along with the physical connections between the OSA and the speech, this chapter presents a brief description of state-of-the-art speaker recognition techniques for OSA voice characterization along with the results and their analysis for OSA assessment by using speech features. The findings extracted from this chapter provide information if there actually exist OSA-related acoustic patterns in the speakers' speech to discriminate subjects according to their level of OSA severity.

4.1 Introduction

First hypotheses about the relationship between acoustic patterns and OSA can be found in first studies that aimed for evolutionary change due to acquisition speech along with the presence of craniofacial abnormalities in subjects with a degree of OSA severity ([Guilleminault et al., 1984](#)), that generally cause a longer and more collapsible upper airway (UA). Consequently, abnormal or particular speech features in OSA speakers may be expected from altered structure or function of their UA.

These hypotheses were supported by first studies carried out by [Fox et al. \(1989\)](#) and [Monson and Fox \(1987\)](#), where authors identified speech disorders in sleep apnea speakers by means perceptive descriptors related to articulation, phonation, and resonance. Later, the first acoustic analysis approaches of speech for OSA detection were introduced by [Fiz et al. \(1993\)](#) and ([Robb et al., 1997](#)), where patterns in acoustic spectral features associated with differences in laryngeal behavior between OSA and non-OSA speakers were found. These early contributions

drove recent proposal for using automatic speech processing in OSA detection.

Different approaches have been studied for Spanish ([Fernández-Pozo et al., 2009](#)) and Hebrew ([Goldshtein et al., 2011](#)). Results have been reported for different types of speech (i.e., sustained and/or continuous speech) ([Solé-Casals et al., 2014](#)), different speech features ([Montero-Benavides et al., 2014](#)), and modeling different linguistic units ([Montero-Benavides et al., 2012](#)). Also, speeches recorded from two distinct positions, upright or seated and supine or stretched, has been considered ([Kriboy et al., 2014b](#)).

Following the automatic speech processing approach for OSA assessment, we propose the study of the state-of-the-art speaker recognition techniques to find OSA-related acoustic patterns in speakers such as those ones reported by previous studies. These techniques are the GMM-Supervectors based on GMM modeling of speech utterances ([Campbell et al., 2006a](#)), and the i-vectors based on the Total Variability modeling ([Dehak et al., 2011a](#)). In last years, these techniques have been the state-of-the-art speaker recognition and language identification systems due to the high capability of speaker voice characterization, and robustness against different sources of session variability. To date these techniques have not been applied into speech-based OSA detection approach, even there is no evidence of studies about their application in medical context, in particular in the detection of vocal fold pathologies or speech disorders.

Besides the speaker recognition and language identification tasks, the high capability of speaker voice characterization of these techniques have led the estimation of some speaker characteristics as age ([Bahari et al., 2014](#)), height ([Poorjam et al., 2015b](#)), and BMI ([Lee et al., 2013](#)). These facts suggest that besides the linguistic message, speech signal carry important information about speakers related to physical or physiological characteristics. In a similar vein, the specific characteristics of the upper airway in OSA individuals have led to hypothesize OSA detection by using these techniques.

4.2 Automatic voice characterization

Although this topic is out of the scope of this Thesis, we are expected that a brief introduction to this topic would help to place the understanding of speech processing techniques in OSA context. It is important to point out that this brief description is based on the state-of-the-art speaker recognition ([Bimbot et al., 2004](#)).

4.2.1 Information in the speech signal

The acquisition of speech in human due to evolutionary anatomical changes along with the presence of abnormalities in the upper airway structure in most of the OSA subjects might produce abnormal acoustic patterns in speech from OSA speakers. Other factors such as behavioral factors (socio-economic status, place of birth, etc.) inherent to the speaker, or the environmental factors (channel acquisition, noise sources, etc.) also add information to the speech signal. Therefore, in the context of the automatic speaker recognition, the goal is to take advantage of the different sources of information available in the speech signal, and combining them in the best possible way in order to identify the speaker identity. The information embedded into the speech signal is often divided into spectral and higher levels.

At the **spectral level**, the information is extracted from the spectrum of the speech signal in short-time windows. The spectrum of the speech signal is directly related to the dynamic configuration of the vocal tract. In this Thesis, we expected that abnormal acoustic patterns related to OSA might be found at this level. Most of the studies carried out in OSA assessment by means of speech processing have been performed at this information level such as features used as frequencies and bandwidth formants (Robb et al., 1997; Goldshtein et al., 2011; Montero-Benavides et al., 2016), and Mel-cepstrum coefficients (Zigel et al., 2008; Fernández-Pozo et al., 2009; Goldshtein et al., 2011).

At the **higher level**, the information into speech can be divided into the following sub-levels:

- **Phonotactic level:** the information of speaker identity is embedded in the particular use of the phones and syllables and their realizations. In speech-based OSA detection context, many studies have approached the analysis of linguistic units at spectral levels such as formant analysis over sustained vowels (Robb et al., 1997; Goldshtein et al., 2011; Montero-Benavides et al., 2016) or vowel nasalization (Montero-Benavides et al., 2014).
- **Prosodic level:** the information is the instantaneous energy, intonation, speech rate and unit durations are analyzed. In speech-based OSA detection context, many studies have approached features at this level to capture phonation abnormalities related to OSA. For instance, Harmonic-to-noise ratio (HNR), the segmental signal-to-dysperiodicity ratio (SDRseg), jitter, and shimmer in (Goldshtein et al., 2011; Montero-Benavides et al., 2014).
- **Idiolectal level:** the information of speaker identity relies on the particular use of the words and language, which not only depends on the speaker but in many other sociolinguistic conditions. In OSA context, the information at this level is not useful for searching

abnormal acoustic cues in OSA speech, but it contributes to the inter-speaker variability, making difficult to identify acoustic patterns within OSA population, for instance idiolect differences between speakers.

4.2.2 Session variability in the speech signal

In speaker recognition systems the main cause of performance degradation is the session variability. This type of variability, usually referred as channel variability, is defined as the set of differences between recordings belonging to the same speaker identity. It is caused by numberless factors that go beyond the acquisition channel such as environmental factors (e.g. speech recorded in different places or situations) and the speech style (e.g. formal or informal speech, conversational or interview speech). Other types of variability can also be found in the speech signal, in particular, in OSA speech, such as:

- **Inter-session variability:** it groups the differences between two recordings belonging to the same identity. There are many factors that can cause this type of variability such as channel acquisition, environment noise, speech style, etc. In this Thesis, this type of variability is not within audio samples of Apnea Databases since they were acquired in a controlled way.
- **Intra-session variability:** it groups the differences within the same recording due to changes in the vocal effort of the speaker, a noise produced in some part of a recording, etc. In this Thesis, this type of variability is likely that it is in audio samples of Apnea Database because of expected speech disorders from OSA speakers.
- **Inter-speaker variability:** it refers to the set of differences between recordings belonging to different identities due just to dissimilarities among them. It represents the discriminative information exploited by the speaker recognition systems in order to perform the recognition task. It is known that part of this information is related to speaker characteristics as weight, age, height, etc.
- **Intra-speaker variability:** it is the set of differences between one or several recordings belonging to the same identity just due to changes related to the identity (e.g. age variation, phone variation, speech style, etc). In this Thesis, this type of variability within the audio samples of Apnea Database, even it might affect the performance of the classification models because of during the speech acquisition the patient is asked to pronounce the sentence as many times in order to avoid misreading and any other spurious phenomena.

- **Nuisance variability:** besides the types of variability described above, in a medical context, there are other sources of variability that might be related to effects of comorbid conditions and other forms of acoustic variability that could affect the system performance. In the case of Apnea Databases, there are subjects affected by other diseases or show other symptoms such as Adenoid Hypertrophy, loud snoring, smoking habits, etc. In consequence, they could contribute to variability in the acoustic space.

To overcome the session variability, which affects directly the performance of speaker recognition systems, there are methods that successfully compensate this type of variability. In the state-of-the-art speaker recognition, the most used are Within-class covariance normalization (WCCN), Linear Discriminant Analysis (LDA), and Nuisance Attribute Projection (NAP). In plain words, those session-compensation methods are based on the projection of the data onto a new space where the noise components can be removed. A detailed description of these methods can be found in (Dehak et al., 2011a).

4.3 Acoustic representation of OSA-related sounds

The speech corpus from our database includes read speech of four sentences that were designed to contain specific distinctive sounds to discriminate between healthy and OSA speakers.

4.3.1 Mel-Frequency Cepstral Coefficients extraction

Once we have a set of utterances containing OSA-specific sounds and collected under a controlled recording environment, speech signals were processed at a sampling frequency of 16 kHz to have a precise wide-band representation of all the relevant information in the speech spectrum. As depicted in Figure 4.1, each utterance was analyzed in speech segments (frames) of 20 ms duration with an overlap of 10 ms, each speech frame was multiplied by a Hamming window. At this point, it should be noted that most of the audio recordings from Apnea Databases have silences of an average length of 2 seconds in the beginning and the end of recordings. Thus, in the pre-processing stage, an energy-based Voice Activity Detector (VAD) operating on the window frames in the time domain was used to identify voiced frames and remove silences and voiceless fragments. After removing the undesired fragments on each frame, the spectral envelope is extracted using Mel-frequency cepstral coefficients (MFCCs).

MFCCs proposed in (Bridle and Brown, 1974) are the most extensively used parameters at the spectral level in automatic speaker recognition systems. It provides a spectral envelope representation of speech sounds extensively used in automatic speech and speaker recognition

Figure 4.1: Mel-frequency coefficient extraction procedure.

(Bimbot et al., 2004; Kinnunen and Li, 2010), pathological voice detection (Godino-Llorente et al., 2006), age (Bocklet et al., 2008), height (Pellom and Hansen, 1997), and body mass index (Lee et al., 2013) estimation. MFCCs have also been used in previous research on speech-based OSA detection (Fernández-Pozo et al., 2009; Goldshtein et al., 2011; Montero-Benavides et al., 2012).

From the MFCC representation, the spectrum magnitude of each frame is first obtained as the absolute value of its discrete Fourier transform (DFT). Then a filterbank of triangular filters spaced in a frequency scale based on the human perception system is used to obtain a vector with the log-energies of each filter (see Figure 4.1). Finally, a discrete Cosine transform (DCT) is applied over the log filterbank energies to produce compact de-correlated MFCC coefficients. Moreover, in order to represent the spectral change over the time, MFCCs are extended to their first order (velocity or ΔMFCCs) derivatives. In this Thesis, in each speech frame i the acoustic information is represented by a D -dimensional vector O_i , called observation vector that includes $19\text{MFCCs} + 19\Delta\text{MFCCs}$ parameters, thus $D = 38$. The extraction of MFCCs was performed

by using HTK toolkit¹.

4.3.2 Utterance modeling

Due to the natural variability in speech production, different utterances corresponding to the same sentence will exhibit variable-duration and thus will be presented by a variable-length sequence $\underline{Q}$ of observation vectors:

$$\underline{Q} = [O_1, O_2, O_3, \dots, O_N] \quad (4.1)$$

where O_i is the D -dimensional ($D = 38$) vector at frame i and N is the number of frames, which will be variable due to the natural variability in speech production, this means that the same sentence read by different speakers will have different lengths. Indeed, the variable-duration explains part of the aforementioned session variability (see section 4.2.2) that affects the performance of speaker recognition and language identification systems (Pelecanos et al., 2004). Nevertheless, there are methods that compensate the effects of utterance duration variability such as those used in the Speaker Recognition Evaluations of NIST, where the length of the enrollment and testing utterance present small variation in a single condition. Moreover, there is a wide range of scenarios where the length of the utterance involved in the recognition process may vary (e.g. forensic applications). In this Thesis, since the acquisition of speech recordings follows a controlled procedure, unlike speaker recognition systems, besides inter- or intra- speaker variability that actually exists in within OSA speech, the existence of session variability due to channel factors (inter-session variability) might be less. In addition, the existence of other types of variability (e.g. comorbid conditions) in the speech recordings of OSA speakers must be approached further.

4.3.2.1 GMM approach

Once a sequence of observations vectors $\underline{Q}$ is extracted from utterances of a given speaker, on the basis of the state-of-the-art speaker recognition, the sequence of vectors that corresponds to OSA and non-OSA groups are used to train two GMMs. Then, by means of a decision score function the final class is obtained according to the threshold-decision, as depicted in Figure 4.2

In the speaker recognition task is frequent that the number of speech samples available for speaker modeling is not enough to robustly train a GMM. As a solution to this limitation, Reynolds et al. (2000) proposed the GMM-UBM framework (see section 3.3.5.2). The underlying

¹htk.eng.cam.ac.uk

Figure 4.2: Modular representation of training and testing phases performed in speech-based OSA detection studies by means of the GMM approach (Fernández-Pozo et al., 2009; Goldshtein et al., 2011; Solé-Casals et al., 2014).

idea is to enhance the speaker's GMM from a Universal GMM, called also Universal Background Model (UBM). The UBM is trained with speech samples belonging to different speakers to represent as much as possible the speaker-independent distribution of the feature vectors. Then, by applying Maximum a Posteriori Adaptation (MAP), the speaker's GMM parameters updated from the UBM. A detailed description of this procedure is given in section 3.3.5.2.

According to the problem formulation approach (section 3.3.1), the goal is to design a function that could be able to assess the OSA severity or predict AHI through supervised machine learning models, in particular, Support Vector Machine (SVM) for OSA severity classification and Support Vector Regression (SVR) for Apnea-Hypopnea Index (AHI) prediction. Therefore, it is required that the input data to these models must be non-variable-length, in consequence, the sequence of observations $\underline{Q}$ in expression 4.1 must be mapped into a vector with fixed dimension. In this Dissertation, the non-variable-length speech features were obtained by two speech feature modeling approaches from the state-of-the-art speaker recognition: GMM-Supervectors and i-vectors. Both approaches have been successfully applied to speaker recognition (Dehak et al., 2011a), language recognition (Dehak et al., 2011b), speaker age estimation (Bahari et al., 2014), speaker height estimation (Poorjam et al., 2015b), and accent recognition (Bahari et al., 2013). We think that their success in those challenging tasks is that speech contains significant sources of interfering inter-speaker variability (i.e. speaker weight, height, etc.). Thus, it is reasonable to explore their use for OSA assessment.

4.3.2.2 GMM-Supervector approach

The success of MAP adaptation along with the fact that only GMM means are adapted from the UBM derived in a new form to represent models: the GMM-Supervector or just Supervector (Campbell et al., 2006a). The supervector is formed by stacking D-dimensional means vectors of a K-component adapted GMM into KD-dimensional Gaussian supervector as depicted in Figure 4.3. As the supervector synthesis the information about a given speaker by a KD-dimensional fixed-length vector, it can be considered as a feature vector that is susceptible to be modeled. This concept was first exploited by (Campbell et al., 2006b), where speaker supervectors served as an input of an SVM system.

Figure 4.3: *GMM-Supervector modeling.*

In this Thesis, the resulting fixed-length vector is suitable to be used as input to classification or regression model such as Support Vector Machine (SVM) or Support Vector Regression (SVR) model to classify subjects according to their OSA's severity or predict Apnea-Hypopnea Index (AHI), respectively. As summarized in Table 4.1, in our experiments, the GMM-UBM training, GMM adaptation, and supervector extraction were done using the MSR Identity ToolBox for MatlabTM (Sadjadi et al., 2013).

In the same manner, as shown in Table 4.1, in order to have a precise acoustic representation for each sentence a GMM with 512 components was used, resulting in a high-dimensional supervector of size 19456. Since the UBM training requires a considerable amount of data so as to represent a global acoustic space, we used several large databases containing microphonic speech sampled at 16 kHz, covering a wide range of phonetic variability from continuous Spanish

speech. A detailed description of these databases is given in section 3.2.3.

Table 4.1: *Description of parameters used by functions in order to extract the speech features (MFCCs, Supervector, i-vector) and train SVM and SVR models for OSA classification and predict AHI.*

Function description	Features	Tool
MFCCs coefficients extraction	<ul style="list-style-type: none"> - #DFT bins = 512 - #filters = 26 - #MFCC coeff. = 19 - #$\Delta\Delta$MFCC coeff. = 19 	HTK toolkit
GMM-UBM training	<ul style="list-style-type: none"> - #GMM comp. = 512 - # expectation maximization iterations = 10 	
GMM-MAP adaptation	- Mean adaptation MAP relevance factor = 10	
Supervector extraction	<ul style="list-style-type: none"> - Stacking GMM means into a single vector - Dim. = #MFCC coeff. $\times$ #GMM comp. Dim. = $38 \times 512 = 19456$ 	MSR Identity Toolbox
Total Variability subspace matrix	<ul style="list-style-type: none"> - Dim. of Total Variability matrix = [400; 300; 200; 100; 50; 30] - # iterations = 5 	
i-vector extraction	<ul style="list-style-type: none"> - Dim. of Total Variability matrix = [400; 300; 200; 100; 50; 30] 	
Support Vector Machine training	Grid search parameters: <ul style="list-style-type: none"> - C, model complexity = $-20 : 20$ - γ, Gaussian kernel parameter = $-2^7 : 2^7$ 	LIBSVM
Support Vector Regression training	Grid search parameters: <ul style="list-style-type: none"> - C, model complexity = $-20 : 20$ - e, insensitive-zone = [0; 0.01; 0.1; 1] - γ, Gaussian kernel parameter = $-2^7 : 2^7$ 	

¹ Hidden Markov Model Toolkit and Speech recognition toolkit. <http://htk.eng.cam.ac.uk/>

² Matlab Toolbox for Speaker Recognition Research. <https://www.microsoft.com/en-us/download/details.aspx?id=52279>

³ A Library for Support Vector Machines. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

As shown in Table 3.6, the whole development dataset includes 25517 speech recordings (~ 33.33 hrs.) from 970 male speakers and 19422 speech recordings (~ 25.43 hrs.) from 659 female speakers. It should be noted that among the development dataset a set of speakers certainly diagnosed with OSA, for both gender, and not used for tests, are included to reflect the OSA-specific characteristics of speech.

According to (González-Domínguez, 2011), until the development of the supervectors the compensation of session variability in the speaker recognition and language recognition topic had been applied on the following two principles:

- Suppress the session variability.
- Treat the variability as a combination of discrete sources rather than continuous.

Based on these principles many methods have been developed in the state-of-the-art speaker recognition and language identification, which most of them successfully outperformed the compensation of session variability. Beyond their relative success, these compensation techniques had limits related to the way of dealing the session variability as categorizing the session conditions or suppress the undesired variability rather than modeling the specific variability within a given recording, since these facts only could be applied into a controllable speech-acquisition environment. These facts suggested the approach of modeling the session variability by means Joint Factor Analysis (JFA), which breaks the two principles described above and states the hypothesis that much of the variability associated with a given recording lies within subspaces of a much lower dimensionality than the original space (González-Domínguez, 2011).

4.3.2.3 I-vector approach

Based on the success of high-dimensional GMM-Supervectors, a new paradigm called i-vector has been successfully and widely used by the speaker recognition community (Dehak et al., 2011a). This approach consists in modeling a single subspace rather than the speaker and session subspaces such as performed in the Joint Factor Analysis approach (Dehak, 2009). This new subspace is called as the Total Variability and contains the speaker and session variability (Dehak et al., 2011a). This new approach is motivated by the empirical findings found in (Dehak, 2009), where besides the session information modeled by the session factors, speaker information is also contained in session factors. Therefore, given an utterance u belonging to the speaker s , the speaker and session-dependent GMM supervector M_{su} can be written as follows:

$$M_{su} = m + Tw_{su} \quad (4.2)$$

where m is the GMM-Supervector of the UBM (speaker- and session-independent), and T defines the Total Variability space (speaker-independent), which is rectangular matrix of low rank representing the primary directions of total acoustic variability observed in a large

development corpora, and w_{su} is a random vector having a normal distribution $N(0, I)$, also called total factor.

The main idea relies on projecting the high dimensional GMM supervector M_{su} into a low dimensional subspace (Total Variability) such that the sum of the prior and the supervector estimation error is minimized. Therefore, the lower error estimation of the prior and the supervector, the higher amount of data is required. As the UBM and Total Variability space represent a global acoustic space and capture the primary directions of total acoustic variability, respectively, their training requires a considerable amount of data. In this Thesis, the UBM and Total Variability space were trained by using the development dataset described in section 3.2.3.

As regard Total Variability matrix T and total factor w_{su} , they are estimated by means of the Expectation-Maximization (EM) algorithm. In the case of total variability matrix estimation, the speaker's utterance recordings used for development are regarded as having been produced by different speakers. In the case of the total factor w_{su} , also called *i-vector*, its estimation comes from the definition of the Baum-Welch statistic extracted from the UBM.

For a given utterance represented by a sequence of observations $\underline{O} = [O_1, O_2, O_3, \dots, O_N]$, the Baum-Welch statistic extracted from the GMM-UBM λ_{UBM} composed of K components are obtained by:

$$n_i = \sum_{t=1}^N P(C_i | o_t, \lambda_{UBM}) \quad (4.3)$$

$$f_i = \sum_{t=1}^N P(C_i | o_t, \lambda_{UBM}) o_t \quad (4.4)$$

where C_i is the i component of the GMM-UBM, and n_i and f_i are the 0_{th} and 1_{st} order statistic, respectively. This procedure is similar to the GMM-MAP means adaptation described in section 3.3.5.2. The estimated GMM supervector components, also called centralized first-order Baum-Welch statistics $\tilde{f}_i$, are given by the equation,

$$\tilde{f}_i = \sum_{t=1}^N P(C_i | o_t, \lambda_{UBM}) (o_t - m_i) \quad (4.5)$$

where m_i is the i component of the GMM-UBM. Therefore, the i-vector for a given utterance w_{su} can be obtained using the following equation:

$$w_{su} = (I + T' \Sigma^{-1} N(u) T)^{-1} T' \Sigma^{-1} \tilde{f}_i(u) \quad (4.6)$$

where $\tilde{f}_i(u)$ is the GMM supervector obtained by concatenating all the first-order Baum-Welch statistic $\tilde{f}_i$ for a given utterance u , T is the total variability matrix, and Σ is a diagonal covariance matrix which models the residual variability not captured by the total variability matrix T . An efficient procedure for training T and w_{su} can be found in (Kenny et al., 2008).

Contrasting both approaches, the i-vector approach has the advantage of projecting the high dimensional supervectors onto a low-dimensional subspace, where most of the speaker-specific variability is modeled.

In this Thesis, the i-vectors w_{su} and the GMM-Supervectors were employed as input to Support Vector Machine (SVM) and Support Vector Regression (SVR) model, as performed in previous studies for speaker's age (Bahari et al., 2014; Shivakumar et al., 2014) and height (Poorjam et al., 2015b) prediction using i-vectors as well speaker's age prediction (Dobry et al., 2011) using GMM-Supervectors. Hereafter the GMM-Supervectors are referred to as supervectors.

4.4 OSA assessment by speech features

4.4.1 Speech features validation

As mentioned before, the i-vectors and supervectors have been used in previous studies for predicting speaker characteristics as age (Bahari et al., 2014; Bahari and Van-Hamme, 2012; Dobry et al., 2011; Shivakumar et al., 2014), weight (Poorjam et al., 2015a), and height (Poorjam et al., 2015b). In order to validate the extraction procedure of our speech features, we compare our results of age, weight and height prediction to those studies. As shown in Table 4.2, some prediction models are those based on high-dimensional supervectors, afterward dimension reduction techniques are applied over them such as Weighted Pairwise PCA in (Dobry et al., 2011) and Weighted Supervised Non-Negative Matrix Factorization in (Bahari and Van-Hamme, 2012). Besides the dimensionality reduction, these techniques aim to eliminate the inter-session speaker variability, similar to most used intersession compensation techniques in the state-of-the-art speaker recognition (Dehak et al., 2011a), such as Whitin-Class Covariance Normalization (WCCN), Linear Discriminant Analysis (LDA), and Nuisance Attribute Projection (NAP).

Contrasting our results and the reference studies, similar performances were achieved for i-vectors and supervectors (Table 4.2).

In the case of i-vector approach for predicting age (Bahari et al., 2014; Shivakumar et al., 2014), height (Poorjam et al., 2015b) and weight (Poorjam et al., 2015a), those studies applied inter-session compensation techniques, in particular, Whitin-Class Covariance Normalization (WCCN). Then, the compensated feature vectors were used as input to prediction model. Results

Table 4.2: *Performance comparison among prediction models for male speakers' age estimation*

Prediction model	MAE (years)	Correlation coefficient
i-vector-WCCN (Bahari et al., 2014)	6.08	0.77 ^a
Supervised i-vector-WCCN (Shivakumar et al., 2014)	8.12	0.48 ^b
Supervector-WSNMF (Bahari and Van-Hamme, 2012)	7.9	-
Supervector-WPPCA (Dobry et al., 2011)	7.6	0.74 ^a
i-vectors (our study)	7.45	0.66 ^b
Supervectors (our study)	7.47	0.65 ^b

^a Level of significance is not reported.^b Correlations are significant beyond the 0.001 level of confidence.

show modest correlations between predicted and true clinical variables values at Fowler scale (see section 3.3.4.4). Despite the inter-session compensation, there are cases where the compensated-model' performance is lower than the degraded respect to no-compensated model due to the large intra-class variation caused by speaker and channel and lower inter-class variation. Therefore, more data might be required in order to emerge the discriminant space directions related to clinical variable groups.

As regards age prediction, in the case of i-vectors, as shown in Table 4.2, our performance is slightly worse than (Bahari et al., 2014) but similar to the others. A plausible explanation of this low performance might be due to that most of the population in (Bahari et al., 2014; Shivakumar et al., 2014) are young people, ranging from 20 to 30 years old whereas most of our OSA speakers are above 45 years old. In the case of supervectors, a similar performance to the others is obtained. According to Cerrato et al. (2000), speech records from young speakers can be better discriminated than those from older ones.

Table 4.3: *Performance comparison among prediction models for male speakers' height estimation.*

Prediction model	MAE (cm)	Correlation coefficient
i-vector-LSSVR (Poorjam et al., 2015b)	6.2	0.40 ^a
Supervector-SVR, Linear kernel (our study)	5.17	0.43 ^b
i-vector-SVR, Linear kernel (our study)	5.0	0.47 ^b

^a Level of significance is not reported.^b Correlations are significant beyond the 0.001 level of confidence.

As regards height prediction (Table 4.3), our estimations performances (both in terms of MAE and correlation coefficient) are comparable and better when using i-vectors, that those in (Poorjam et al., 2015b). As regards weight prediction (Table 4.4), in (Poorjam et al., 2015a) a novel approach was proposed based on fusion at the feature level. Our estimation performances are only slightly worse than Poorjam et al. (2015a). Although its weight-population distribution is similar to ours, the amount of training data available could explain its good performance.

Table 4.4: *Performance comparison among prediction models for male speakers' weight estimation.*

Prediction model	MAE (kg)	Correlation coefficient
i-vector-NFA-LSSVR (Poorjam et al., 2015a)	11.16	0.56 ^a
Supervector-SVR, Linear kernel (our study)	11.73	0.40 ^b
i-vector-SVR, Linear kernel (our study)	11.59	0.41 ^b

^a Level of significance is not reported.

^b Correlations are significant beyond the 0.001 level of confidence.

For the rest of clinical variables, previous research has also demonstrated moderate results when estimating speakers' cervical perimeter from speech such as (Evans et al., 2006) and less success when estimating body mass index (Hamdan et al., 2013). Only positive results have been reported in (Lee et al., 2013), although they have questioned the possibility of overfitting in their experiments as they used feature selection methods over a large set of acoustic features.

In summary, the results are very similar to those published by other studies, which is a good indicator of the validity of features.

4.4.2 Apnea-Hypopnea Index prediction

As described in section 3.3.1.1, the regression method aims to design an estimator function f for the Apnea-Hypopnea Index (AHI) as well as the estimation of other clinical variables for validating results, such that, the difference between the estimated and the truth value is minimized.

In the following tables, prediction results using supervectors and i-vectors for all our clinical variables are listed on Tables 4.6-4.8. It should be noted that the prediction results using i-vectors were obtained by testing both linear and RBF kernels, but only the linear kernel was considered for supervectors because of their large dimensionality makes it not advisable mapping them into a higher dimensional space since the nonlinear mapping does not improve performance obtained by using linear mapping (Chang and Lin, 2011).

Results in Table 4.5 indicate that most of the prediction results are more accurate than the prior estimator. Comparing the AHI prediction results in Tables 4.6-4.8, both in terms of MAE and Pearson correlation, i-vectors of dimension 100 (MAE = 13.63; $r = 0.30$; p -value < 0.001; 95% CI [0.21, 0.37]) is slightly better than supervectors (MAE = 13.94; $r = 0.25$; p -value < 0.001; 95% CI [0.17, 0.33]), both by using a SVR-linear kernel. As regards the clinical variables prediction, most of best results were also obtained by using low dimensionality i-vectors (i.e. 200, 100) with SVR-linear kernel. In contrast, comparing these results with those reported for predicting clinical variables such as age (Bahari et al., 2014) and height (Poorjam et al., 2015b),

the best prediction results were obtained by using a high dimensionality i-vector (400).

The dimensionality for either the Joint Factor Analysis (JFA) or Total Variability is related to the number of factors in their respective subspaces; in JFA the speaker and channel factors, whereas in Total Variability the identity factor (i.e. i-vectors). For instance, for the speaker verification task, in (Dehak et al., 2011a), the JFA and the i-vector approach were tested. The former used 300 speaker factors and 100 channel factors, whereas in the latter for training the Total Variability subspace they used the sum of total factors: 400 factors. In addition, Dehak et al. (2011a) tested different session compensation techniques such as Within-Class Covariance Normalization (WCCN) and Linear Discriminant Analysis (LDA). The best performance was obtained by i-vectors 400 with WCCN compensation.

Table 4.5: *Male speakers' AHI and clinical variables estimation by the prior estimator.*

Variable	Mean Absolute Error (MAE)
AHI (event/h)	14.49
Height (cm)	5.72
Age (years)	8.12
Weight (kg)	12.82
BMI (kg/m ²)	3.82
CP (cm)	2.50

Abbreviation: BMI, Body Mass Index; CP, cervical perimeter.

Beyond that, this fact might suggest that there is no a rule or established method for selecting the dimensionality of i-vectors (i.e. the number of total factors). Similar to the JFA approach, the number of speaker and channel factors are selected according to their values of eigenvalues associated with the eigenvoices and eigenchannels, respectively. In the same manner, this approach can be applied for selecting the number of identity factors (i.e. i-vector dimensionality) for predicting AHI.

Table 4.6: *Male speakers' AHI and clinical variables estimation using supervectors and SVR (linear kernel).*

Variable	Mean Absolute Error (MAE)	Correlation Coefficient
AHI (event/h)	13.94	0.25
Height (cm)	5.17	0.43
Age (years)	7.47	0.65
Weight (kg)	11.73	0.40
BMI (kg/m ²)	3.64	0.27
CP (cm)	2.16	0.50

Abbreviation: BMI, Body Mass Index; CP, cervical perimeter.

Note: All correlation coefficients are significant beyond the 0.001 level of confidence.

Table 4.7: *Male speakers' AHI and clinical variables estimation using i-vectors and SVR (linear kernel).*

Variable	Mean Absolute Error (MAE)						Correlation coefficient					
	i-vector dimension						i-vector dimension					
	400	300	200	100	50	30	400	300	200	100	50	30
AHI (event/h)	14.05	14.21	13.98	13.63	13.94	14.05	0.21	0.22	0.22	0.30	0.27	0.22
Height (cm)	5.10	5.20	5.15	5.00	5.26	5.30	0.45	0.41	0.42	0.48	0.36	0.37
Age (years)	7.61	7.59	7.45	8.12	8.42	8.82	0.65	0.65	0.66	0.59	0.53	0.45
Weight (kg)	12.07	12.09	11.96	11.59	12.05	12.03	0.33	0.32	0.36	0.41	0.37	0.34
BMI (kg/m ²)	3.57	3.55	3.57	3.64	3.69	3.70	0.29	0.33	0.32	0.30	0.26	0.22
CP (cm)	2.21	2.18	2.17	2.23	2.29	2.35	0.45	0.48	0.49	0.46	0.42	0.35

Abbreviation: BMI, body mass index (AHI); CP, cervical perimeter.

Note: All correlation coefficients are significant beyond the 0.001 level of confidence.

Table 4.8: *Male speakers' AHI and clinical variables estimation using i-vector and SVR (RBF kernel).*

Variable	Mean Absolute Error (MAE)						Correlation coefficient					
	i-vector dimension						i-vector dimension					
	400	300	200	100	50	30	400	300	200	100	50	30
AHI (event/h)	13.95	13.80	13.78	13.65	13.98	14.24	0.23	0.28	0.24	0.28	0.23	0.19
Height (cm)	5.05	5.12	5.02	5.09	5.26	5.31	0.46	0.43	0.46	0.46	0.38	0.38
Age (years)	7.50	7.51	7.72	7.83	8.35	9.08	0.66	0.66	0.63	0.62	0.55	0.40
Weight (kg)	11.84	11.78	11.73	11.83	12.31	13.10	0.39	0.40	0.42	0.40	0.31	0.19
BMI (kg/m ²)	3.57	3.58	3.58	3.71	3.92	3.94	0.30	0.28	0.34	0.25	0.11	0.09
CP (cm)	2.20	2.13	2.17	2.29	2.42	2.52	0.46	0.51	0.48	0.43	0.33	0.21

Abbreviation: BMI, body mass index (AHI); CP, cervical perimeter.

Note: All correlation coefficients are significant beyond the 0.001 level of confidence.

Following the same approach, Figure 4.4 (left) shows the MAE evolution for AHI prediction using i-vectors with different size, ranging from 30 to 400 where the minimum MAE is achieved by i-vectors of dimension 100 for both kernels. From that point the MAE tends to increase. The main reason for this behavior can be attributed to additional directions captured beside the main direction that probably contains the information related to OSA. This “optimal” point is the inflection point where the change in tendency occurs. This point can be also detected by inspecting the eigenvalues associated with each i-vector. As depicted in Figure 4.4 (right), the elbow in the curve, at the point 100, gives an insight of the appropriate number of i-vectors. Therefore, the higher the associated eigenvalue the higher the confidence to represent the main OSA directions, whereas i-vectors with or nearby zero associated eigenvalue should be discarded.

To state the main directions captured by i-vectors are related to OSA information might lead to misunderstanding since, besides the possible captured OSA-variability by the i-vector, it is known the high capability of i-vectors to represent other clinical variables such as age and height

(Tables 4.7 and 4.8). Therefore, these findings related to AHI prediction might be associated with the indirect association to those clinical variables, which will result in false discoveries. This behavior is also applied to supervectors since they are also good predictors for clinical variables estimation. These aspects will be discussed further on.

Figure 4.4: a) Evolution of the MAE for male speakers' AHI prediction in function of the numbers of i-vectors considered. b) Eigenvalues associated with the i-vectors estimated sorted in descending order.

As reported in Tables 4.6 - 4.8, poor results for AHI prediction were obtained using both i-vectors and supervectors. Best performance was obtained by i-vectors of dimension 100 (MAE = 13.63; $r = 0.30$; p -value < 0.001; 95% CI [0.21, 0.37]). In terms of strength of the correlation and according to (Fowler et al., 1998), this correlation is *weak* (i.e. correlation coefficients between 0.2 and 0.39). This poor correlation can be seen through the regression plot in Figure 4.5 (left), where there is a trend towards to plateau at AHI > 30 (severe OSA group), which indicates a bias to underestimate AHI in that group. In contrast, for patients considered as non-mild and non-severe OSA (AHI < 10) there is a bias to overestimate the AHI.

Examining the the agreement between two methods (AHI_{EST} vs. AHI_{PSG}) through the Bland-Altman plot in Figure 4.5 (right), there are a large limits of agreement (-29 to 41), where 95% of cases of the differences between two methods lie between these limits. This large limits indicates a lack of agreement between both methods where the prediction model (AHI_{EST}) may be 29 events/h below or 41 events/h above the reference model (AHI_{PSG}). In the same manner, this lack of agreement can be also seen through the mean values of AHI_{EST} and AHI_{PSG} values, where those above 30 events/h are underestimated (i.e. dots above the mean difference line 6.0 (red line) in Figure 4.5 (right)), whereas those values below 10 are overestimate (i.e. blue dots below the mean difference line 6.0). Nevertheless, some values within 10 and 30 lie close to the mean difference line 6.0.

Figure 4.5: a) Plot regression for male speakers' AHI estimation by means *i*-vectors of dimension 100. b) Bland-Altman plot for male speakers' AHI estimation results by means *i*-vectors of dimension 100. Points inside the rectangular area (black line) correspond to the percentage of diagnostic agreement between the two methods.

In order to get a detailed visualization of the percentage of AHI values that lie within the true OSA severity range, as well as, underestimated and overestimated values, in the following tables these rates are shown.

Table 4.9: Percentage of estimated and overestimated AHI values of the non-OSA ($AHI < 10$) male population using *i*-vectors and supervectors.

Feature	$AHI_{est} < 10$ (estimated)	$AHI_{est} \geq 10$ (over-estimated)
Supervectors	49 (30.62%)	111 (69.38%)
<i>i</i> -vector(dim=400)	32 (20.00%)	128 (80.00%)
<i>i</i> -vector(dim=300)	36 (22.50%)	124 (77.50%)
<i>i</i> -vector(dim=200)	35 (21.87%)	125 (78.13%)
<i>i</i> -vector(dim=100)	46 (28.75%)	114 (71.25%)
<i>i</i> -vector(dim=50)	46 (28.75%)	114 (71.25%)
<i>i</i> -vector(dim=30)	38 (23.75%)	122 (76.25%)

As shown in Tables 4.9 and 4.11, there are high percentage of over- and under- estimated AHI values that belongs to the non-OSA ($AHI < 10$) and the severe-OSA ($AHI > 30$) groups, respectively. This is related to the poor agreement between methods in those AHI ranges. Contrasting both extreme-severe groups, in the severe OSA group, a higher percentage of estimated AHI values are under-estimated. Indeed, this is associated with there is a trend towards to plateau at $AHI > 30$ as shown in 4.5 (left).

Moreover, as shown in Table 4.10, more than 70% of predicted AHI values are within the mild-OSA range ($10 \leq AHI_{est} \leq 30$). These findings could suggest the good agreement between

Table 4.10: *Percentage of estimated, underestimated and overestimated AHI values of the mild-OSA ($10 \leq AHI \leq 30$) male population using i-vectors and supervectors.*

Feature	$AHI_{est} < 10$ (under-estimated)	$10 \leq AHI_{est} \leq 30$ (estimated)	$AHI_{est} > 30$ (over-estimated)
Supervectors	22 (10.19%)	166 (76.85%)	28 (12.96%)
i-vector(dim=400)	8 (3.70%)	198 (93.98%)	5 (2.32%)
i-vector(dim=300)	19 (8.80%)	189 (87.50%)	8 (3.70%)
i-vector(dim=200)	22 (10.19%)	184 (85.19%)	10 (4.62%)
i-vector(dim=100)	28 (12.96%)	179 (82.87%)	9 (4.17%)
i-vector(dim=50)	25 (11.57%)	189 (87.50%)	2 (0.93%)
i-vector(dim=30)	17 (7.87%)	198 (91.66%)	1 (0.47%)

Table 4.11: *Percentage of estimated and underestimated AHI values of the severe OSA ($AHI > 30$) male population using i-vectors and supervectors.*

Feature	$AHI_{est} \leq 30$ (under-estimated)	$AHI_{est} > 30$ (estimated)
Supervectors	126 (84.56%)	23 (15.44%)
i-vector(dim=400)	148 (99.33%)	1 (0.67%)
i-vector(dim=300)	146 (97.98%)	3 (2.02%)
i-vector(dim=200)	141 (94.63%)	8 (5.37%)
i-vector(dim=100)	135 (90.60%)	14 (9.40%)
i-vector(dim=50)	143 (95.97%)	6 (4.03%)
i-vector(dim=30)	140 (93.96%)	9 (6.04%)

methods in that OSA group. Nevertheless, it does not make sense since a good performance could be expected in the opposite effect, that is, best discrimination between extreme-severe groups such as non-OSA ($AHI < 10$) and severe OSA ($AHI > 30$). This “good agreement” might be due to the prediction model trends to predict all samples as the mean, similar to the prior model (Table 4.5). Accordingly, these findings provide some considerations about the incorrect approach of similar machine learning pitfalls could lead to false discoveries or misinterpretations of results.

To measure the degree of agreement between both methods, we use the Diagnostic Agreement Approach (see section 3.3.4.3). In Table 4.12 the Diagnostic Agreement Approach errors for predicting AHI using i-vectors a supervectors are shown.

As can be seen in Table 4.12, the percentage of agreement between methods is around 50% for both i-vectors and supervectors, although there is no a cut-off of percentage related to a good percentage of agreement, this agreement is below of that reported in (Kriboy et al., 2014a), where the reported agreement between methods was 66.7%. In Figure 4.5 (right), the diagnostic agreement samples are surrounded by a black line rectangular area. From those samples, more than 50% corresponds to Mild OSA group (see 4.13). This finding suggests that the model shows

Table 4.12: *Diagnostic agreement approach for male speakers' AHI estimation using i-vectors and supervectors.*

Feature	Agreement	Over-estimation	Under-estimation	No-agreement
Supervectors	48.76%	17.90%	16.00%	17.34%
i-vector(dim=400)	50.48%	13.52%	19.24%	16.76%
i-vector(dim=300)	50.86%	14.10%	18.48%	16.56%
i-vector(dim=200)	49.90%	15.24%	17.71%	17.15%
i-vector(dim=100)	52.95%	13.52%	16.57%	16.96%
i-vector(dim=50)	49.33%	13.14%	20.76%	16.77%
i-vector(dim=30)	48.76%	14.48%	20.19%	16.57%

a bias to predict all samples as the mean.

Table 4.13: *Diagnostic Agreement percentage for each OSA severity group using i-vectors of dimension 100, and supervectors.*

Feature	Agreement percentage		
	Control (AHI < 10)	Mild OSA ($10 \leq \text{AHI} \leq 30$)	Severe OSA (AHI > 30)
Supervectors	40.00%	54.86%	5.14%
i-vector(dim=100)	39.56%	54.68%	5.76%

These findings provide a better understanding of the trend towards to plateaus at $\text{AHI} > 30$ depicted in Figure 4.5 (left), Although Kriboy et al. (2014a) reported good performances of AHI prediction ($r = 0.67$, $p\text{-value} < 0.01$, $\text{DA} = 66.7\%$), their input vectors for training and testing were composed of speech features and clinical variables, in particular, BMI and age information, which are already known as good predictors of AHI due to their relationship with OSA risk factors. Accordingly, these results might be related to false discoveries. This aspect will be reviewed in Chapter 5.

In brief, the above results indicate the poor performance for AHI estimation by means of speech features. Moreover, in terms of agreement, there is a lack of agreement between both methods: the prediction model and the polysomnography (PSG). Despite there is a high percentage of diagnostic agreement with the PSG for predicting AHI values in the mild OSA group ($10 \leq \text{AHI} \leq 30$), there is a suspicion regression model tends to predict test samples as the average of all training samples, as performed by the prior estimator (Table 4.5).

In addition, it could be useful to evaluate these speech features in other conditions that might contribute either to session variability or false discoveries due to confounder factors. Consequently, in the following, the resulting prediction models are compared to the best prediction model based on the best i-vector configuration, which is the i-vector of dimension 100. It is im-

portant to point out that, hereafter, this i-vector configuration will be referred to as the original model.

4.4.2.1 Population without differences in clinical variables

It is known that clinical variables are good predictors of AHI (Hoffstein and Szalai, 1993) as well as speech features (i.e. i-vectors and supervectors) are good predictors of clinical variables, as shown in Tables 4.6 - 4.8, thus it could explain, in terms of correlation, the performance of AHI prediction using speech features, in particular, the best i-vector configuration: dimension 100 ($r = 0.30$; p -value < 0.001 ; 95% CI [0.21, 0.37]). This idea is supported when there are differences, in terms of clinical variables, between OSA (AHI ≥ 10) and non-OSA (AHI < 10) groups, in particular, significant differences in terms of age, weight, BMI, and cervical perimeter. Consequently, the degradation of performance for AHI prediction is expected when these differences in clinical variables between both groups are removed from the whole population under study. To assess the impact of clinical variables on speech features performance, we selected subjects with no differences in the clinical variables, in particular on body mass index ($26 \leq \text{BMI} \leq 31$), and age ($33 \leq \text{age} \leq 44$).

Table 4.14: Contrast analysis among Control (AHI < 10) and OSA (AHI ≥ 10) groups on a subject without differences on body mass index (BMI) and age including mean and standard deviation (mean $\pm$ SD).

Variable	Control (n=19)	OSA (n=24)	p-value
Age (years)	41.26 $\pm$ 1.59	40.66 $\pm$ 1.43	NS
Weight (kg)	90.15 $\pm$ 9.96	87.37 $\pm$ 8.52	NS
Height (cm)	177.42 $\pm$ 8.70	175.79 $\pm$ 5.74	NS
Body mass index (kg/m ²)	28.57 $\pm$ 1.48	28.21 $\pm$ 1.50	NS
Cervical perimeter (cm)	41.26 $\pm$ 2.86	41.46 $\pm$ 1.63	NS
AHI (event/h)	6.18 $\pm$ 2.76	28.82 $\pm$ 13.02	< 0.001

Abbreviation: NS, nonsignificant.

As can be seen in Table 4.14, once the possible effect of clinical variables in speech is minimized, few samples are remained ($n = 43$). To evaluate the i-vectors under these new conditions, we followed the regression approach, which is training a prediction model to predict AHI. However, it is clear that one of the inconvenient is the small population of subjects to properly train a prediction model, in particular, a Support Vector Regression (SVR) model. Besides a degradation of prediction performance is expected due to the effect of clinical variables is removed in both Control and OSA groups, the small number of samples and high dimensionality of speech features is also related to its performance degradation.

It is important to point out that the number of speech samples used for the evaluation does not correspond to the number of subjects, but it is the number of utterances per subject (see Table 3.4 and 3.5). The number of utterances used for this experiments are shown in Table 4.15.

Table 4.15: *Number of utterances for control and OSA groups in uniform male population in terms of each type of sentence.*

Sentence	All	Control	Mild OSA	Severe OSA
1	59	20	25	14
2	71	24	28	19
3	59	20	25	14
4	71	24	28	19
All sentences	260	88		172

In order to evaluate the speech features, the experiments were repeated on the new population and results are shown in following table.

Table 4.16: *AHI prediction in a male population without differences in clinical variables by using i-vectors of dimension 100.*

Prediction model	Mean Absolute Error (MAE)	Correlation Coefficient	p-value
Prior	11.85	-	-
i-vectors100	12.05	0.22	NS

Abbreviation: NS, nonsignificant.

Results in Table 4.16 indicate that there is a noteworthy degradation in performance. Besides the clear effect of removing the differences in clinical variables from the population, there are other factors related to machine learning pitfalls that might be associated with this degradation. Among these factors, there are: small number of samples and high dimensionality features for training. For this experiment, the feature dimensionality is 100, and the number of utterances samples is 260, in consequence, there is a high chance of overfitting, nevertheless, due to the complexity of the prediction model (SVR) and the validation scheme followed in this Thesis, this effect is minimized, hence the low performance could be related to the small number of samples for training and testing. This effect was also reported in our previous studies (Espinoza-Cuadros et al., 2015, 2016). In (Espinoza-Cuadros et al., 2015), the experiments were performed on the Apnea v2.0 database, the database comprised 285 male subjects and the best performance was obtained by using low dimensionality i-vectors ($r = 0.1$, $p\text{-value} < 0.001$). One year later, in (Espinoza-Cuadros et al., 2016), the number of samples in the Apnea v2.0 database increased, and it contained 426 male subjects. The best performance was also obtained by using low dimensionality i-vector ($r = 0.3$, $p\text{-value} < 0.001$). Thus, the number of samples used for training plays an important role in the model performance.

4.4.2.2 The Apnea Databases' sentences

In speaker recognition systems the main cause of degradation performance is due to the session variability, in particular, inter-session variability (section 4.2.2) defined as the differences between recordings belonging to the same speaker identity. In the context of this Thesis, the model performance could be affected by this type of session variability since different types of sentences were recorded for each subject, as described in section 3.2.1.2. To evaluate this effect, we performed the AHI prediction for each type of sentence. The number of utterances of the dataset used for this experiment is described in Table 3.5. For this experiment, we used the best i-vector configuration: dimension 100.

Table 4.17: *Male speakers' AHI estimation using i-vectors of dimension 100 for each type of sentence.*

Type of sentence	Mean Absolute Error (MAE)	Correlation Coefficient
1	14.00	0.21
2	13.90	0.27
3	14.02	0.21
4	14.23	0.24

Abbreviation: NS, nonsignificant.

Note: all correlation coefficients are significant beyond the 0.01 level of confidence.

As shown in Table 4.17, despite the session variability was removed there is a slight performance degradation as compared to the original model (see Table 4.7). It suggests that inter-session variability could affect the model to a less extent. Nevertheless, the small number of samples for training might also contribute to this degradation. In consequence, the real impact of the inter-session variability cannot be assessed under these conditions.

From the above results (Table 4.17), best performances were obtained for sentences 2 and 4. These high performances might be related to the information of possible abnormalities carried on these sentences such abnormal nasalization and phonation due to fatigue in the upper airway muscle respiratory. But, this finding cannot be confirmed since the small number of samples contributes to the degradation of performance, in consequence, the “real” performance of this approach cannot be evaluated.

4.4.2.3 Session compensation analysis

As described in section 4.2.2, the most common session variability compensation techniques are Within-class covariance normalization (WCCN), Linear Discriminant Analysis (LDA), and Nuisance Attribute Projection (NAP). Besides the successful performances achieved by these techniques in the state-of-the-art speaker recognition systems, they also involve less computa-

tional costs in low dimensionality (e.g. i-vectors) features as compared to high dimensionality features (e.g. supervectors).

These techniques have also been tested for other tasks such as age prediction (Bahari et al., 2014). Two different approaches based on Within-class covariance normalization (WCCN) were tested and compared to the age prediction model. The approaches are: 1) age variability compensation, 2) inter-session variability compensation. From results, the age variability compensation did not perform well, in contrast, the inter-session variability compensation outperformed the prediction model. Despite the age compensation model was trained by using age labeled groups, the large intra-class variability caused by speaker, channel, etc differences results in significant overlap of the age classes. Thus, Bahari et al. (2014) conclude:

“... with smaller class differences, more data is required before the discriminant sub-space directions emerge from the estimation noise, even more so under a model describing classes with only second order.”

Based on (Bahari et al., 2014), we proposed two different approaches for session compensation for AHI prediction: 1) OSA variability compensation, and 2) inter-session variability compensation. For this experiment, we followed the validation scheme described in section 3.3.3.1. In addition, besides the searching for the prediction models’ optimal hyperparameters, the training data in each fold was used for training the WCCN matrix. From results, both approaches achieved poor performances as compared to the original model, in terms of correlation, no significant correlation between true and predicted AHI values were obtained. Similar to previous results, the small number of samples has a great influence on the final performance, which makes difficult to assess the real performance of it and the degradation of the performance is expected. Nevertheless, as pointed by Bahari et al. (2014), these poor performances might also be influenced by the smaller differences between classes. Thus, more data is required to represent the main directions of variability.

In summary, the above results reaffirm the hypothesis about the poor characterization capability of the speech features for OSA speakers. Despite the success of the state-of-the-art speech characterization techniques for speaker or language recognition at spectral level (see section 4.2.1), it seems that there are no clear insights for OSA speaker characterization. Although the non-cepstral and supra-segmental information are not carried by i-vectors (González-Rodríguez et al., 2014), which could explain their poor performance, other types of non-spectral and supra-segmental related features such as hyponasality, speech rate, fundamental frequency, pitch range phonation have been also tested and good results were obtained (Montero-Benavides et al., 2014;

Solé-Casals et al., 2014), nevertheless, there are some machine learning pitfalls that lead to misleading results (Foster et al., 2014).

4.4.3 OSA severity classification

One of the main issues addressed by classification models is the imbalanced classes as described in section 3.3.3. Its incorrect treatment is still one of the main pitfalls of studies based on machine learning topics, in particular, biomedical engineering research as pointed out by Foster et al. (2014). In the context of speech-based OSA detection, a straightforward and less thoroughness approach for OSA severity classification could be the classification of test samples through their estimated AHI values by means a prediction model. Nevertheless, considering the imbalanced classes, this approach will clearly prone to classify the samples as the majority class. In this Dissertation, we managed this issue by training a classification model in order to overcome the imbalanced classes problem. This issue along other machine-related issues were discussed in detail in Chapter 5.

As described in section 3.3.1.2, the classification approach aims to train a classifier such that for an unseeing subject the detection error could be the minimum along with a high discriminatory and generalization performance. In our research, we employed the SVM model, which overcome the imbalance classes issue by means the optimization of its weighted cost function (3.3.2.1). Similar to regression results, OSA classification results using i-vectors were obtained by testing both linear and RBF kernels, whereas only linear kernel was considered for supervectors.

Table 4.18: *OSA severity classification for male speakers using the SVM model (linear kernel).*

Feature	Acc(%)	Sen(%)	Spec(%)	DOR(95% CI)	AUC(95% CI)
Supervectors	62.85	70.13	46.25	2.02 [1.37, 2.96]	0.63 [0.60, 0.65]
i-vector(dim=400)	64.38	70.13	51.25	2.46 [1.68, 3.61]	0.63 [0.58, 0.68]
i-vector(dim=300)	64.95	70.13	53.12	2.66 [1.81, 3.90]	0.65 [0.60, 0.70]
i-vector(dim=200)	64.38	70.13	51.25	2.46 [1.68, 3.61]	0.68 [0.62, 0.72]
i-vector(dim=100)	66.10	70.13	56.87	3.09 [2.10, 4.55]	0.69 [0.64, 0.70]
i-vector(dim=50)	64.95	70.13	53.12	2.66 [1.81, 3.90]	0.65 [0.60, 0.70]
i-vector(dim=30)	63.42	70.13	48.12	2.17 [1.48, 3.19]	0.62 [0.57, 0.67]

Results in Tables 4.18 and 4.19 indicate no remarkable differences in classification performance in terms of type of kernel. Nevertheless, the highest performance is obtained by the linear kernel and using i-vectors of dimension 100 (Table 4.18). In terms of classification rates (Sen, Spec, DOR) and AUC, the best performance is obtained using i-vectors of dimension 100 (Acc = 66.10%; Sen = 70.13%; Spec = 56.87%; AUC = 0.69; 95% CI [0.64, 0.70]). It is important to point out that the classification rates are reported by optimal thresholds, which are

Table 4.19: *OSA severity classification for male speakers using the SVM model (RBF kernel).*

Feature	Acc(%)	Sen(%)	Spec(%)	DOR(95% CI)	AUC(95% CI)
i-vector(dim=400)	64.76	70.13	52.50	2.59 [1.77, 3.80]	0.65 [0.59, 0.70]
i-vector(dim=300)	64.38	70.13	51.25	2.46 [1.68, 3.61]	0.65 [0.59, 0.70]
i-vector(dim=200)	66.66	70.13	58.75	3.34 [2.27, 4.92]	0.67 [0.62, 0.73]
i-vector(dim=100)	65.90	70.13	56.25	3.01 [2.05, 4.34]	0.68 [0.63, 0.72]
i-vector(dim=50)	64.57	70.13	51.87	2.53 [1.72, 3.71]	0.67 [0.61, 0.71]
i-vector(dim=30)	61.52	70.13	41.87	1.69 [1.15, 2.48]	0.61 [0.56, 0.66]

determined by searching the maximum sensibility under the constraint of a minimum percentage of specificity. This criterion is described in the section 3.3.4.2.

Class prevalence is also a condition that clearly plays a role in the final performance of a classifier. That means that the probability that we get a patient that actually suffers from OSA or a healthy subject varies on the conditions of the experiment or how we interpret the prevalence of this disease (i.e. prior probabilities). We could take as reference the current prevalence of the syndrome of the world male population, then the OSA prevalence should be in between 3% to 7% (Lam et al., 2010), but this prevalence could change if we constraint the analysis to a specific population. Thus, the application of similar conditions in different experimental contexts could not perform as expected.

4.4.3.1 OSA extreme cases classification

As pointed out by the epidemiologic studies, there is a group of symptoms reported by most of OSA diagnosed subjects such as heavy snoring, witnessed apneas, and daytime hypersomnolence, where the level of symptoms' manifestation relies on the OSA severity in subject (Lam et al., 2010). Therefore, the symptoms of severe OSA diagnosed subjects ($AHI > 30$) are manifested to a great extent. In the case of speech abnormalities, those might be expected to be stronger as compared to mild or low OSA severity. This fact was reported in (Solé-Casals et al., 2014), where they performed classification experiments for OSA assessment contrasting different severity levels: 1) “extreme cases” that involves non-OSA ($AHI < 5$) and severe-OSA cases ($AHI > 30$), and 2) “in-between cases” that involves OSA($AHI \geq 15$) and non-OSA($AHI < 15$) cases. Best performance in (Solé-Casals et al., 2014) was obtained in “extreme cases”, whereas there was a drop of performance when the classification was performed “in-between cases”.

Following a similar approach to (Solé-Casals et al., 2014), we evaluated the speech features in OSA “extreme cases” and OSA “in-between cases”. Our extreme cases are the severe OSA patients ($AHI > 30$) and non-OSA patients ($AHI < 10$), as shown in Table 4.20, whereas the “in-

Table 4.20: Contrast among Control ($AHI < 10$) and severe OSA ($AHI > 30$) groups including mean and standard deviation ($mean \pm SD$).

Variable	Control (n=160)	OSA (n=149)	p-value
Age (years)	44.22 ± 11.47	50.48 ± 11.64	< 0.001
Weight (kg)	84.26 ± 12.68	98.87 ± 20.25	< 0.001
Height (cm)	175.33 ± 7.70	175.38 ± 7.03	NS
Body mass index (kg/m^2)	27.41 ± 3.73	32.06 ± 5.86	< 0.001
Cervical perimeter (cm)	40.55 ± 2.79	43.47 ± 3.28	< 0.001
AHI	4.86 ± 2.94	46.64 ± 14.07	< 0.001

Abbreviation: NS, nonsignificant.

between cases” involves the non-OSA ($AHI < 10$) and OSA ($AHI \geq 10$) cases. It is important to point out that the “in-between cases” correspond to the classes evaluated in section 4.4.3, whose results are shown in Tables 4.18 and 4.19. For this experiment, we used the best i-vectors configuration obtained for OSA severity classification (Table 4.18) that is the i-vectors of dimension 100.

Table 4.21: OSA extreme and in-between cases classification for male speakers using SVM model (linear kernel) and i-vectors of dimension 100

OSA severity cases	Acc(%)	Sen(%)	Spec(%)	DOR(95% CI)	AUC(95% CI)
extreme cases	66.34	70.46	62.50	3.97 [2.47-6.40]	0.71 [0.65, 0.77]
in-between cases	66.10	70.13	56.87	3.09 [2.10, 4.55]	0.69 [0.64, 0.70]

As shown in Table 4.21, the OSA severity classification in “extreme cases” obtained slight better results than the “in-between cases”. This result agrees with Solé-Casals et al. (2014), where best classification rates were also obtained in OSA severity extreme cases. Nevertheless, the influence of clinical variables might also be the cause of the high performance in OSA “extreme cases” classification. As shown in Table 4.20, there are significant differences between Control and severe OSA groups except for height, as a result the influence of clinical variables into i-vectors for OSA assessment is expected. Despite the influence of clinical variables into speech features could be removed by filtering subjects with significant differences between both groups, the small number of samples and high dimensionality features resulting in a high chance of falling on machine-learning pitfalls make difficult the assessment of the classification model.

4.5 Discussion

From regression and classification results presented above, poor performances were achieved. In the case of regression approach, the best result was obtained using i-vectors of dimension 100

(MAE = 13.63; $r = 0.30$; p -value < 0.001; 95% CI [0.21, 0.37]). In terms of diagnostic agreement approach, this i-vector configuration also reached the best rate of agreement between methods (DA = 52.95%). These poor performances is also related to regression model's bias for under- or over-estimating the AHI values within OSA severity groups: AHI < 10 (overestimate) and AHI > 30 (underestimate) as well as the regression model is prone to predict in-between OSA severity ($10 \leq \text{AHI} \leq 30$) samples as the average of the training data values (e.g. prior model). Moreover, it is important to point out the influence of the clinical variables on speech features for AHI prediction, where the age and body mass index (BMI) are correlated with AHI as well as the i-vectors are also good predictors of those clinical variables (see Table 4.7). Despite the effect of the clinical variables was removed by filtering samples under a criterion, the resulting small number of samples as compared to i-vector dimensionality was not enough to properly train a prediction model, as a result low performance was obtained (Table 4.16). In the same manner, the OSA-sentences-related variability was removed by training a model using each type of sentence, nevertheless, low performances were achieved as compared to experiments with the whole population. The causes of performances' degradation might be the small number of samples for training as well as the poor characterization of speech features for each sentence. Besides OSA-sentences-related variability, there are others that also contribute to performances' degradation: the inter-session variability. Based on (Bahari et al., 2014), two different approaches for session compensation were carried out based on the Within Class Covariance Normalization (WCCN): 1) OSA variability compensation, and 2) inter-session variability compensation. From results, no improvements in results were achieved as compared to the original model. As regards classification approach, low performances were achieved. Best result was obtained using i-vectors of dimension 100 (Acc = 66.10%; Sen = 70.13%; Spec = 56.87; DOR = 3.09; AUC = 0.69; 95% CI [0.64, 0.70]). In the case of OSA "extreme cases", there is a slight improvements as compared to the "in-between" OSA severity cases. According to (Solé-Casals et al., 2014), this behavior might be due to the changes in upper airway structure, where its closure may be so pronounced in OSA severity group, thus voice is much affected.

At this point, some questions could arise about the type of speech characterization used for OSA assessment, in particular, spectral characteristics such as the Mel-Frequency Cepstral coefficients, and their representation by GMM modeling as supervectors and i-vectors. As described in section 4.3.2, a speaker's utterance modeling comes from the adaptation of a Universal Background Model (UBM) previously trained with a considerable amount of data in order to represent a global acoustic space to the specific acoustic characteristic of speakers. In the same manner, the i-vectors as a representation of supervectors in low dimensionality by representing

the primary directions of global acoustic space are extracted from the Total Variability subspace, which is also trained with a large amount of data. In this Thesis, the data for training the UBM and Total Variability matrix was Spanish utterances of speakers, all samples recorded following the same protocol and recording conditions similar to Apnea databases, and most of them were asymptomatic individuals ($\sim 98\%$), whereas the remainder show OSA-related symptoms. It is important to point out that those subjects were not used either for training or testing, as described in section 3.2.3. Therefore, it is evident that the low performance of both supervectors and i-vectors might be due to OSA voices were not used for training the UBM in order to represent the possible OSA acoustic space, in particular, i-vectors, where the primary directions extracted from the Total Variability subspace are related to the speaker variability instead of OSA-related acoustic information. Before to discuss this issue, it is important to point out the high amount of data required for training the UBM and Total Variability subspace. The number of hours of recordings in the state-of-the-art speaker recognition system is above 400 hours (Burg et al., 2008), whereas our development dataset has only 35 hours of recordings. Therefore, the insufficient number of samples along with the asymptomatic condition of speakers from the development dataset might have been contributed to the poor performance of speech features. In order to evaluate this hypothesis we have trained the Total Variability matrix using only the Apnea Database v2.0 such that the OSA-related acoustic variability had been captured. In Tables 4.22 and 4.23, the AHI prediction and OSA severity classification results are presented.

Table 4.22: *Male speakers' AHI estimation using i-vectors and SVR (linear kernel). The development data used for the Total Variability subspace training is the Apnea Database v2.0.*

Feature	Mean Absolute Error (MAE)	Correlation Coefficient(95% CI)
i-vector(dim=400)	13.90	0.25 [0.16, 0.32]
i-vector(dim=300)	13.85	0.23 [0.14, 0.30]
i-vector(dim=200)	14.11	0.24 [0.15, 0.31]
i-vector(dim=100)	13.82	0.29 [0.21, 0.36]
i-vector(dim=50)	14.10	0.26 [0.17, 0.33]
i-vector(dim=30)	14.17	0.22 [0.13, 0.29]

Note: All correlation coefficients are significant beyond the .001 level of confidence.

According to results in Tables 4.22 and 4.23, there is a drop of performance respect to the original models (see Tables 4.7 and 4.18). Despite optimistic results were expected due to the training data was also used as development data, which could have contributed to model overfitting, these poor performances might indicate there is an insufficient number of samples to properly represent the acoustic variability in OSA speakers within the Total Variability subspace.

In brief, our OSA speech characterization approach might be limited by the modeling ap-

Table 4.23: *OSA severity classification for male speakers using SVM model (linear kernel). The development data used for the Total Variability subspace training is the Apnea Database v2.0.*

Feature	Acc(%)	Sen(%)	Spec(%)	DOR(95% CI)	AUC(95% CI)
i-vector(dim=400)	65.90	70.13	56.25	3.01 [2.05, 4.43]	0.67 [0.61, 0.71]
i-vector(dim=300)	64.57	70.13	51.87	2.53 [1.72, 3.71]	0.64 [0.59, 0.69]
i-vector(dim=200)	64.76	70.13	52.50	2.59 [1.77, 3.80]	0.64 [0.59, 0.69]
i-vector(dim=100)	65.90	70.13	56.25	3.01 [2.05, 4.43]	0.68 [0.63, 0.73]
i-vector(dim=50)	64.95	70.13	53.12	2.66 [1.81, 3.90]	0.66 [0.61, 0.71]
i-vector(dim=30)	63.61	70.13	48.75	2.23 [1.52, 3.27]	0.64 [0.58, 0.69]

proach performed (i-vectors, supervectors), the confounder variables such as the clinical due to the relationship with the OSA-risk factors (aging, overweight), and other variability sources as well as the inter-session variability. The search for other types of speech characterization techniques or speech modeling approaches with techniques to compensate unwanted sources of variability could shed light on this issue. Nevertheless, considering some recent studies as (González-Rodríguez et al., 2014), we should also acknowledge the limitation of i-vectors to represent relevant segmental (non-cepstral) and supra-segmental (voice quality and prosodic features) speaker information. Therefore, subspace projection techniques could also be explored over other speech acoustic features previously related to OSA as: nasality (Goldshtein et al., 2011; Fernández-Pozo et al., 2009), voice turbulence (Solé-Casals et al., 2014; Murillo-Blanco et al., 2013) or specific co-articulation trajectories. In the same line, a comparative analysis of results for both different recording positions, as proposed in (Kriboy et al., 2014b), should also be addressed.

Other factors could also contribute to the variability in the speech such as the limitation of recording speech at night when patients go to sleep unit. Other acoustic differences could be expected at different times of the day, for instance, soon after waking in the morning voice could reflect laryngeal trauma from overnight snoring. In fact, recent research has already pointed at other possible effects on patient voices that can help in OSA detection such as analyzing speech recorded from a supine or stretched position (Kriboy et al., 2014b; Solé-Casals et al., 2014), and acoustic analysis of nose and mouth breath sounds recorded in supine and upright position (Premkumar et al., 2014; Solà-Soler et al., 2014). Our results are also limited to the speech from Spanish speakers, in particular, Andalusian dialect (Málaga), as compared to other languages will require a thorough analysis of language-dependent acoustic traits in OSA voices.

Chapter 5

Review of speech-based OSA detection models

In the previous Chapter, our results when estimating age, height and weight are comparable to those previously published in (Bahari et al., 2014), (Poorjam et al., 2015b), and (Poorjam et al., 2015a). However, OSA prediction results contrast markedly with those reported in previous research connecting speech and OSA, which reached good performances. Among these possible machine learning issues we have considered two ones that could be related to good performances in previous research:

1. The use of prediction models derived from speech features, which are also correlated with other patient characteristics (age, weight, etc) that act as confounding factors.
2. The methodology and validation method followed, in particular, the wrong address of machine learning pitfalls (e.g. high dimensionality, small training dataset, feature selection). This lack of thoroughness in the methodology and validation method in some of these studies is reflected as not many information is provided.

Therefore, in this Chapter, we present a critical review of previous studies that led us to identify possible machine learning issues similar to those reported in (Foster et al., 2014) as well as the analysis of the clinical variables influence in the OSA prediction.

5.1 Clinical variables as OSA predictors

It is known the relation between the OSA and some clinical variables, in particular, those related to risk factors such as the overweight and aging. The linear relationship between the OSA and

clinical variables can be observed through the Pearson correlation between clinical variables and the Apnea-Hypopnea Index. As shown in Table 5.1, correlations between overweight-related clinical variables (weight, BMI, cervical perimeter) and AHI can be described as a “modest” correlation ($0.40 \leq r \leq 0.69$) according to (Fowler et al., 1998). This modest correlation can be seen through the OSA assessment by means of a prediction or a classification model using the clinical variables as predictors.

Table 5.1: *Pearson’s correlation between clinical variables and AHI on the male population.*

	AHI	Height	Age	Weight	BMI	CP
AHI	1					
Height	-	1				
Age	0.15	-0.34	1			
Weight	0.40	0.39	-0.12	1		
BMI	0.42	-	-	0.89	1	
CP	0.41	0.13	0.15	0.72	0.72	1

Abbreviation: BMI, body mass index; CP, cervical perimeter.

Notes: For clarity, non-significant correlation values were omitted. Correlation coefficients are significant beyond the .001 level of confidence.

Table 5.2: *Male speakers’ AHI estimation using clinical variables.*

Method	MAE	Correlation Coefficient	Agreement rate (%)
SVR linear kernel	12.71	0.44	54.67
SVR RBF kernel	13.43	0.38	50.10

Notes: All correlation coefficients are significant beyond the 0.01 level of confidence.

As shown in Tables 5.2 and 5.3, best performances were obtained using linear kernel. As regards AHI prediction, contrasting Tables 5.2 and 4.7, better results were obtained ($r = 0.44$; p -value < 0.001 ; 95% CI [0.36, 0.50]) using clinical variables than i-vectors ($r = 0.30$; p -value < 0.001 ; 95% CI [0.21, 0.37]). Those findings confirm clinical variables as good predictors for OSA. As shown in Figure 5.1 (left), these good performances can be also observed through the less trend to plateau at AHI values above 30 event/h as compared to speech features in Figure 4.5 (left). Nevertheless, as described in the previous chapter, a good correlation does not imply a good agreement between both prediction methods. The Bland-Altman plot in Figure 4.5 (right)

Table 5.3: *OSA classification severity for male speakers using clinical variables.*

Method	Acc(%)	Sen(%)	Spec(%)	DOR(95% CI)	AUC(95% CI)
SVM linear kernel	68.76	70.13	65.62	4.48 [3.01, 6.66]	0.74 [0.69, 0.78]
SVM RBF kernel	68.00	70.13	63.12	4.02 [2.71, 5.94]	0.72 [0.67, 0.76]

shows more narrow limits of agreement as compared to speech features 4.5 (right). By means of the diagnostic agreement approach (Table 5.2), the good prediction performance does not correspond with the by diagnostic agreement rate ($DA = 54.67\%$). Despite there is a slight improvement as compared to the i-vector ($DA = 52.95\%$), those percentages of agreement are below than those reported in (Kriboy et al., 2014a) using speech features ($DA = 66.70\%$). In consequence, the clinical-variables-based OSA prediction model and PSG test cannot be used interchangeably.

Figure 5.1: a) Plot regression for male speakers' AHI estimation using clinical variables. b) Bland-Altman plot for male speakers' AHI estimation results using clinical variables. Points inside the rectangular area (black line) are the percentage of diagnostic agreement between the two methods.

The good performance of clinical variables for OSA detection is also displayed in the classification results, as shown in Table 5.3. Comparing performances of clinical-variables and speech features, more balanced classification rates, in terms of sensibility and specificity, were obtained using clinical variables ($Acc = 68.76\%$; $Sen = 70.13\%$; $Spec = 65.62\%$) as compared to i-vectors of dimension 100 ($Acc = 66.10\%$; $Sen = 70.13\%$; $Spec = 56.87\%$). In the same manner, better AUC is obtained by clinical variables ($AUC = 0.74$; 95% CI $[0.69, 0.78]$) than i-vectors ($AUC = 0.69$; 95% CI $[0.64, 0.70]$).

The above results confirm the clinical variables as good predictors of OSA, and considering also that speech features are good predictors of clinical variables, as shown in Table 4.7, a considerable influence over the prediction or classification model for OSA assessment using speech features is expected, consequently, both models could be prone to false discoveries.

Table 5.4: *Test characteristics of previous research using speech analysis and supervised models for OSA severity classification and AHI prediction.*

Study	Population description		Classification			Regression	
	Control	OSA	Acc.(%)	Sen.(%)	Spec.(%)	AUC	r
Fernández-Pozo et al. (2009)	40 M (AHI < 10)	40 M (AHI > 30)	81	77.5	85	-	-
Montero-Benavides et al. (2012)	40 M (AHI < 10)	40 M (AHI > 30)	85	-	-	-	-
Solé-Casals et al. (2014)	79 W+48 M (AHI ≤ 5)	20 W+101 M (AHI ≥ 30)	82.85	81.49	84.69	-	-
Goldshtein et al. (2011)	14 W (AHI ≤ 5)	19 W (AHI > 5)	-	86	84	-	-
	12 M (AHI ≤ 10)	48 M (AHI > 10)	-	83	79	-	-
Kriboy et al. (2013)	25 M (AHI ≤ 10)	78 M (AHI > 10)	80	80.65	80	0.85	-
Kriboy et al. (2014a)	131 M ¹		-	-	-	-	0.67 ²
i-vectors approach (our study)	160 M (AHI < 10)	365 M (AHI ≥ 10)	66.10	70.13	56.87	0.69	0.30
	108 W (AHI < 10)	96 W (AHI ≥ 10)	68.62	70.83	66.66	0.73	0.28

¹ population description is not provided in (Kriboy et al., 2014a).² results using speech features plus age and BMI.

Abbreviation: W, women; M, men.

5.2 OSA and speech connection

From results of Chapter, our results in Table 4.7 are comparable to those published for age (Bahari et al., 2014), height (Poorjam et al., 2015b), and weight (Poorjam et al., 2015a) prediction. However, in the case of OSA assessment, our results contrast markedly with previous research. Therefore, in this section, a critical review of previous studies (including ours) has been addressed to identify possible machine learning issues as well as differences between our research and the seminal studies of the connection between speech and OSA. It is important to point that part of this review was presented in (Espinoza-Cuadros et al., 2016). In Table 5.4, the most relevant existing research proposals using automatic speech analysis and machine learning for OSA assessment are presented.

Therefore, machine-learning issues along with differences between these studies and others are described below.

5.2.1 Differences in formants frequencies and bandwidths

A first discrepancy, not related to machine learning issue, was addressed in (Montero-Benavides et al., 2016), where notable differences were found with respect to the seminal work by Robb et al. (1997). In (Robb et al., 1997), statistically significant differences between OSA and non-OSA speakers were found for several formants frequencies and bandwidths extracted from sustained vowels, in contrast, Montero-Benavides et al. (2016), only revealed very weak correlations with two formant bandwidths. These differences might be attributed to the small and biased samples used in exploratory analysis in (Robb et al., 1997), that is 10 male subjects diagnosed with OSA,

where most of the diagnosed patients present severe OSA ($AHI > 30$), and 10 non-OSA male subjects, while in (Montero-Benavides et al., 2016), they explored a larger sample of 241 male subjects, all diagnosed with OSA and showing a wide range of AHI values.

5.2.2 Database size and patient's clinical condition

In (Fernández-Pozo et al., 2009) and (Montero-Benavides et al., 2012), speech samples from Control ($AHI < 10$) and OSA ($AHI > 30$) subjects were used to train a binary machine learning classifier for severe OSA detection. Healthy and OSA speakers were thus classified using two models: one trained to represent OSA voices and the other to model healthy voices. Two different approaches were performed: 1) a text-independent approach using two Gaussian Mixture Models (Fernández-Pozo et al., 2009), and 2) through two text-dependent Hidden Markov Models (Montero-Benavides et al., 2012). Correct classification rates were 80% and 85% for Gaussian Mixture Models and Hidden Markov Models, respectively. However, these very good results contrast with ours, where a weak relationship between speech and AHI and low OSA severity classification performance were found (see Table 4.18). Consequently, we replicated the experiments in (Fernández-Pozo et al., 2009) and (Montero-Benavides et al., 2012) on the Apnea Database v2.0, and found a significant degradation of performance achieving correct classification rates of 68% and 67% for Gaussian Mixture Models and Hidden Markov Models, respectively. This important reduction in performance is due to the very small database. In (Fernández-Pozo et al., 2009) and (Montero-Benavides et al., 2012) the database used was the Apnea Database v1.0 (section 3.2.1), which includes 40 asymptomatic and patients with low OSA ($AHI < 10$), and 42 OSA subjects with severe OSA ($AHI > 30$), while in this Dissertation, the experiments were performed on the Apnea Database v2.0, which includes 160 subjects with low OSA ($AHI < 10$), mild-OSA subjects ($10 \leq AHI \leq 30$) and 149 OSA severe subjects ($AHI > 30$). As pointed out in (Foster et al., 2014), the size of training and evaluation sets are important factors to gain a reasonable understanding of the performance of any classifier. Furthermore, another relevant factor that could explain this degradation in performance is that in Apnea Database v1.0 the 40 part of control subjects in Fernández-Pozo et al. (2009) and (Montero-Benavides et al., 2012) were asymptomatic individuals and selected trying to have both control and OSA population as matched as possible in terms of age and BMI, while in the Apnea Database v2.0 all individuals are suspected to suffer from OSA as they have been referred to a sleep disorders unit, and most of them present common clinical symptoms as heavy snoring, witnessed apneas and daytime sleepiness.

Following the same approach previously described in (Fernández-Pozo et al., 2009), the

connection between OSA and speech analysis were also studied for Hebrew language ([Goldshtein et al., 2011](#); [Kriboy et al., 2013](#)). The study in ([Goldshtein et al., 2011](#)) trained two GMMs to classify between OSA and non-OSA speakers by using short- and long-term spectral features such as MFCC, LPC, etc. Overall 100 features were extracted for each frame (i.e. short-term) and 28 features were extracted from several of the subject's frame (i.e. long-term). Previous to the GMM modeling, a set of features were selected by means of a feature selection method. The experimental protocol presented by [Goldshtein et al. \(2011\)](#) properly separated female and male speakers, and different AHI thresholds were used to define OSA and non-OSA groups: an AHI threshold of 5 was used for women and 10 for men. Results reported a specificity of 83% and sensitivity of 79% for OSA detection in men and a specificity of 86% and 84% for females. The major limitation in this study is the small number of cases under study: a total number of 60 male subjects (12 controls vs. 48 OSA), and 33 female subjects (14 control vs. 19 OSA).

Besides the low reliability with such small samples, a critical issue in ([Goldshtein et al., 2011](#)) is the use of feature selection techniques over a large number of features (sometimes on the order of hundreds) and small number of samples. Moreover, the population under study showed significant differences between the OSA and non-OSA groups, in particular, the OSA diagnosed subjects were on average ten years older and more overweight. Considering these differences, they performed an age- and BMI- matched experiments, and results reported a slight drop in performance compared to the previous approach. Later, the same research group reported the same behavior in ([Kriboy et al., 2013](#)), but over a different database with 103 male subjects. [Kriboy et al. \(2013\)](#) pointed out that mismatch could be explained by the use of a different database with more subjects and with different balance in terms of possible confounding factors such as BMI, age, etc.

5.2.3 Confounder variables for OSA detection

Another possible cause of overoptimistic results in ([Fernández-Pozo et al., 2009](#)) and ([Montero-Benavides et al., 2012](#)) can be traced considering possible indirect influences of speech and AHI mediated through the clinical variables, as described by means of the correlation between AHI and the clinical variables (Table 5.1) and clinical variables predictions using speech features, as presented in Table 4.7.

As discussed in ([Goldshtein et al., 2011](#)) speech features can be less correlated with AHI than with some clinical variables as age or BMI that are good predictors of AHI ([Hoffstein and Szalai, 1993](#)). Therefore, a population of Control and OSA speakers with significant differences in these confounding variables can be prone to false discovery of discrimination results due to

the underlying differences in these confounders variables but not in AHI. This fact was reported in (Montero-Benavides et al., 2014), where good results for OSA detection by using 16 speech features, which many of these features are similar to those traditionally used in detecting voice pathologies, such as HNR, Jitter, Shimmer, and others. The good performance was obtained by testing the model in the training dataset that had significant differences between both Control and OSA groups in terms of age and BMI. Nevertheless, the performance was degraded when their model was tested on a dataset designed to avoid statistically significant differences in age and BMI.

5.2.4 Feature selection and gender-independent

Besides exploring and reporting on significant differences in confounding speaker’s features such as age, BMI, height, etc., the same criterion must be extended to any other factor that could also affect speech such as speakers’ dialect, gender, mood state, and so forth. We believe that this issue could explain the good discrimination results when detecting severe OSA reported in (Solé-Casals et al., 2014), where they analyzes both sustained and connected speech and recordings from two distinct positions: upright or seated and supine or stretched. The reason for recording from two distinct uttering positions, which was also preliminarily explored in (Kriboy et al., 2014b), is that due to anatomical and functional abnormalities in OSA individuals, different body positions can affect differently their vocal tract, thus presenting more discriminative acoustic features. In addition, Solé-Casals et al. (2014) evaluated several feature selection, feature combination (i.e. PCA) and classification schemes (i.e. Bayesian classifiers, KNN, Support Vector Machines, Neural Networks, and Adaboost), and best results are achieved when using a genetic algorithm for feature selection. An interesting result in (Solé-Casals et al., 2014) is that positive classification results, all above 80%, were obtained when classifying between extreme OSA severity cases: that is OSA severe ($AHI \geq 30$) and non-OSA ($AHI \leq 5$). Nevertheless, a notable reduction in performance was obtained when classifying “in-between cases”, which are mild OSA subjects ($5 \leq AHI \leq 30$) and a AHI threshold of 15 is the border between non-OSA and OSA. This weak estimation was also reported in our results for the whole AHI range using speech features, where a slight increase in performance was achieved in OSA extreme cases classification (Table 4.21) as compared to the whole AHI range (Table 4.18).

Therefore, two critical issues can be identified in (Solé-Casals et al., 2014). First, feature selection is applied over high dimensionality features (i.e. 253) compared to the number of cases (i.e. 248) Even though, authors report the use of cross-validation for the development and evaluation of different classification algorithms there is no clear indication on what data was

Table 5.5: *Wilcoxon two sampled test for MEAN_HNR_VA_A contrasting gender and group of extreme OSA male speakers.*

Gender contrast			Extreme OSA male speakers contrast		
Female ($n = 171$)	Male ($n = 426$)	p -value	Control ($AHI \geq 5$, $n = 69$)	OSA ($AHI \leq 30$, $n = 129$)	p -value
19.43 ± 3.98	17.07 ± 4.23	$< .001$	17.46 ± 3.89	16.38 ± 4.32	0.06

used for feature selection. As discussed in (Foster et al., 2014; Smialowski et al., 2010), feature selection can lead to over-fitted results based on a set of “ad-hoc” selected features.

A second highly relevant issue in (Solé-Casals et al., 2014) is that when evaluating the classification performance between extreme, OSA and control groups contain very different percentages of male and female speakers: 48 men/79women in control vs. 101men/20 women in OSA. This notable imbalance between female and male percentages in both groups is clearly due to the significantly lower prevalence of OSA in women compared to men (Guilleminault et al., 1988). Consequently, considering the important acoustic differences between female and male voices (Titze, 1989), this makes gender a strong confounding factor that might explain the good discrimination results.

In order to illustrate these issues, we have analyzed the best discriminative feature reported in (Solé-Casals et al., 2014), that is the mean value of the Harmonic to Noise Ratio (HNR) measured for the sustained vowel /a/ recorded in seated position. This feature is referenced as **MEAN_HNR_VA_A** in (Solé-Casals et al., 2014). Since Apnea Database v2.0 also contains speech records of sustained /a/ recorded in seated position for both 426 male individuals and 171 female individuals, we have made Wilcoxon two-sample test for **MEAN_HNR_VA_A** values contrasting: 1) a group of male speakers vs. a group of female speakers, and 2) a group of extreme OSA male speakers. Test analysis results in Table 5.5, clearly reveal that while significant differences (p -value $< .0001$) appear contrasting female and male voices, which has already been reported in other studies such as (Titze, 1989), no significant differences were found between extreme OSA severity groups on male speakers (p -value = 0.06). Therefore, this is an illustrative example of how gender can act as a strong confounding factor in the OSA classification, where good rates of classification were achieved, as presented in Table (5.4).

In the same manner, Kriboy et al. (2014a) used a similar prediction scheme to the scheme presented in this dissertation: GMM supervectors used as input to a Support Vector Regression (SVR) model to estimate the AHI. Nevertheless, differently from our regression approach, feature selection was firstly used to select the most 5 discriminative features from a set of 71 acoustic features, and then GMM means supervectors are trained for that small number of features.

Table 5.6: *Speaker’s AHI estimation using supervectors generated by five high-order cepstral and LPC coefficients. These selected features were found by Kriboy et al. (2014a).*

Features	MAE	Correlation coefficient	<i>p</i> -value
$a_{15}, \Delta\Delta c_9, a_{17}, \Delta\Delta c_{12}, c_{16}$	14.23	0.22	$< .001$
age + BMI	12.97	0.40	$< .001$
$a_{15}, \Delta\Delta c_9, a_{17}, \Delta\Delta c_{12}, c_{16}$ + age + BMI	13.04	0.49	$< .001$

Although the experimental protocol followed in (Kriboy et al., 2014a) separated training and validation data in order to avoid overfitting, the set of selected features was composed by five high-order cepstral and LPC coefficients ($a_{15}, \Delta\Delta c_9, a_{17}, \Delta\Delta c_{12}, c_{16}$) which are difficult to interpret or justify. Both cepstral and LPC coefficients are commonly used to represent the acoustic spectral information in speech signals, but higher order coefficients are generally less informative and noisy (Huang et al., 2001).

Another notable limitation to validate results in (Kriboy et al., 2014a) is that SVR regression is applied after adding two clinical variables, age and BMI, to the speech supervector generated from the five selected features. These two clinical variables are well-known predictors of AHI (Hoffstein and Szalai, 1993). Thus, it should have been advisable first to report AHI estimation results only using supervectors representing speech acoustic features, then to present results only using age and BMI, and finally give results extending supervectors with age and BMI.

Therefore, for the sake of the completeness of this dissertation, we have replicate the same estimation procedure described in (Kriboy et al., 2014a) on our database: Apnea Database v2.0. Results in Table 5.6, from top to bottom, show prediction results for AHI using GMM means supervectors from the set of five selected features, clinical variables as age and BMI, and supervectors of selected acoustic features extended with BMI and age.

As shown in Table 5.6, estimation results are mainly driven by the presence of BMI and age, and very weak correlation is obtained when only the set of 5 selected acoustic features is used. Accordingly, it is reasonable to conclude that the well-known correlation between AHI and BMI and age along with possible overfitting due to feature selection on a high number of features with respect to the number of cases might cause the optimistic results presented in (Kriboy et al., 2014a).

5.3 Summary

This review represents an important and useful example to illustrate the potential pitfalls in the development of machine learning techniques for diagnostic applications, in this case the

speech-based OSA detection. The contradictory results using the state-of-the-art speaker recognition and machine learning for OSA assessment over, to the best of our knowledge, the largest database used in this kind of studies, led us to address a critical review of previous studies reporting positive results in connecting OSA and speech. As described in different studies in biomedical research community, several limitations in the development of machine learning techniques were observed and studied ([Foster et al., 2014](#); [Smialowski et al., 2010](#)). Among the observed deficiencies are: the impact of a small size of training and evaluation datasets in performance evaluation, the likelihood of false discoveries or spurious association due to the presence of confounding variables, and the risk of overfitting when the features selection techniques are applied over the high dimensionality features and small training set.

Chapter 6

Craniofacial phenotype characterization in Obstructive Sleep Apnea

THIS CHAPTER PRESENTS AND DETAILS the craniofacial characterization techniques based on photographic analysis ([Lee et al., 2009a,b, 2010](#)), object's appearance modeling ([Cootes et al., 2001](#)), and deep learning-based face recognition techniques.

6.1 Introduction

The basis for OSA appearance is associated with anatomic and functional abnormalities of the upper airway (UA). These changes are related to craniofacial changes ([Cakirer et al., 2001](#); [Stradling and Crosby, 1991](#)) as well as an excess of regional adipose tissue, and overweight ([Young et al., 2005](#)). Nevertheless, the craniofacial morphology is also recognized as an important interacting factor in OSA pathogenesis ([Lee et al., 2009a](#)).

Craniofacial differences between OSA and control subjects have been mainly examined using sophisticated image analysis on magnetic resonance imaging (MRI) such as cephalometry, mainly focused on bony structures ([Guilleminault et al., 1984](#); [Lowe et al., 1995](#)), and soft tissue structure of the upper airway ([Lowe et al., 1995](#); [Lyberg et al., 1989](#); [Schwab et al., 2003](#)). Despite the high reliability and accuracy of these image analysis techniques, they are limited to research applications due to their expense, time-consuming analysis and invasive to the patient.

As an alternative to these methods, a qualitative photographic analysis technique based on anthropometry and photogrammetry were proposed in [Lee et al. \(2009a\)](#) for craniofacial

assessment. This technique aims to search for craniofacial phenotype differences between OSA and non-OSA subjects using a range of measurements (distances, areas, angles) related to face, mandible, eyes, nose, head, and neck. Among these craniofacial measurements, those related to face showed significant differences between OSA and non-OSA subjects such as the width of the face, which was significantly greater in OSA subjects. Later, these significant differences in face-related measurements in OSA population were confirmed in (Lee et al., 2010), where a close relationship between face-related measurements and abnormalities in the upper airway structure in OSA individuals were found out. Therefore, it seems that the craniofacial phenotype of OSA subjects can be described by these features, as stated by Lee et al. (2010):

It is probably that the surface facial phenotype reflects the underlying bony framework, but, in addition, we hypothesized that surface facial dimensions capture phenotypic information that also relates to obesity and upper airway anatomy, both of which are important risk factors for OSA.

Although the anatomic basis for such a relationship between surface facial measurements and OSA is currently unknown, the above hypothesis along with the significant differences in craniofacial measurements between Control (non-OSA) and OSA subjects reported in (Lee et al., 2009a) might disclose a facial phenotype that could certainly describe the OSA population. Actually, that concept is used for OSA diagnosis by physicians, as described in (Cheung et al., 2016), where they carried out a study to determine the proficiency of anesthesiologists, otolaryngologists, and internists at predicting the presence of OSA by visual photographic analysis without the use of a computer program to assist. The motivation behind of the visual inspection performed by the healthcare providers is that the OSA might be associated with grave perioperative complications, thus there is the necessity of a reliable and fast screening method instead of following a routine screening for OSA through questionnaires necessity in the preoperative interview. Among physical characteristics associated with OSA usually visualized by the care providers there are obesity, increased neck circumference, limited neck extension, a high modified Mallampati score, enlarged tongue or tonsils, decreased thyromental distance, and male gender. The accuracy achieved for OSA detection by visual inspection achieved low percentages (i.e. 61 to 64%), but this accuracy was slightly improved with the knowledge of the clinical patient information and their comorbidities.

In brief, in this chapter, we explore the craniofacial characterization techniques as well as alternative measurements based on (Lee et al., 2009a) to represent the characteristic associated with the craniofacial phenotype in OSA subjects. Moreover, the breakthrough of deep

learning-based face recognition techniques opens a new front for the craniofacial phenotype characterization in OSA due to the robustness and reliability to extract information from faces photographs that were captured under less controlled photography capture process. This innovative technique was also described and evaluated in this chapter.

6.2 Review of Automatic Face Analysis

During the past few years, the facial analysis has become one of the most successful applications within image analysis and understanding. The wide range of commercial and surveillance applications, continuous development of feasible technologies such as mobile technologies, and different challenges faced by machine recognition for face recognition have contributed to its advance. In the literature, the problem is usually formulated as follows: given an image or a sequence of images (e.g. video), similar to the speaker recognition approach, the goal is either the identification or verification of persons in a scene using a stored database of faces. In general, automatic face analysis systems involve three stages as depicted in Figure 6.1: face segmentation (i.e. face detection on image), feature extraction, and face recognition (Zhao et al., 2003).

Figure 6.1: Block diagram summarizing basic approach of automatic face recognition system.

One of the most important stages in this system is the facial feature extraction. At this stage, the facial image is represented or encoded in a compact and structured representation of facial characteristics to describe the variability of a particular phenotype from a population or group of individuals. In the literature, many studies in psychology and neuroscience about how human visual perception works in face recognition have been carried out, in particular the findings in psychology about the importance of facial features have been noted in the automatic face recognition literature (Chellappa et al., 1995). There are different approaches related to the facial features categorization, following a comprehensible categorization proposed in (El-Rai et al., 2015), feature extraction methods can be divided into:

- **Holistic methods:** these methods attempt to identify faces using global representations based on the entire facial image as the raw input to a recognition system. Some of the most of widely used representations of the face region are the Principal Components Analysis

(PCA) based Eigenfaces ([Turk and Pentland, 1991](#)) and Linear Discriminant Analysis (LDA) based on Fisherface ([Belhumeur et al., 1997](#)).

- **Local feature based methods:** these methods extract local features from facial regions, for instance, the eyes, nose, and mouth. These features include geometric features, which analyze the distribution of landmarks position (eyes, mouth, nose) and their geometric relationship such as distances, area, angles, etc, and texture features, which describe the appearance of the local patch around each landmark. Appearance-based features include learned images filters such as Gabor filters ([Wiskott et al., 1997](#)), Scale-Invariant Feature Transform ([Bicego et al., 2006](#)), and Local Binary Patterns ([Ahonen et al., 2006](#)).
- **Statistical model-based methods:** these types of feature extraction methods are based on the theory of statistical analysis of shape. One of the most famous approaches is the Active Appearance Model ([Cootes et al., 2001](#)), which is based on learning shape and texture variability of the face across a set of training face images.

In this dissertation, following a similar approach in ([Lee et al., 2009a](#)), a set of local features were proposed. Nevertheless, our major differences as compared to [Lee et al. \(2009a\)](#) are the use of a human-supervised automatic landmarking method and the definition of craniofacial measurements adapted to our less controlled photography capture process.

In addition, many of best performed craniofacial features reported in ([Lee et al., 2010](#)) were surface facial measurements such as the width of the face, binocular width, and cervicomental, which are significantly greater in OSA subjects ([Lee et al., 2009a](#)). Accordingly, the variability of these measurements could be described by their shape variation from the landmarks' position distribution. Therefore, to model the shape variation of facial features, we have tested the Active Appearance Model ([Cootes et al., 2001](#)).

Moreover, successful performances of deep neural networks in computer-vision-related tasks (object recognition, face recognition, scene classification, etc.) have significantly overcome the state of the art in many applications due to the availability of large quantities of training data ([Parkhi et al., 2015](#)), minimizing the proneness to the error as compared to the “handcrafted” features (local features). Nevertheless, recent advances in deep learning have conducted its adaptation to new scenarios related to specific object recognition tasks (e.g. cars' model recognition) and fewer quantities of training data ([Razavian et al., 2014](#)). Therefore, we have also explored the deep learning-based face recognition techniques for craniofacial assessment.

6.2.1 Face Alignment

Before the feature extraction stage, it is necessary to identify first the parts or geometric structures of faces that we want to represent or encoded. The identification of these parts are performed by means of the automatic identification of facial points around or centered at the facial component such as eyes, mouth, and nose. This task is referred to as Face Alignment. In the state-of-the-art face recognition, there are two typical tasks where Face Alignment plays an important role, in both cases the goal is to improve the robustness against pose variation ([Jin and Tan, 2017](#)).

- Face registration: in this case the goal is to locate some major facial point and use them as reference points in order to obtain canonical alignment of the face based on translation, scale, and rotation.
- Feature-based face recognition: this type of face recognition system consist in matching local features from different identities by means of a similarity measure in consequence the performance relies on the face alignment accuracy.

In the literature, there are two approaches that perform the face alignment. On the one hand, early approaches are based on shape parameters estimation of a given image by iteratively adjusting a deformable model, which contains the prior knowledge of face shape and appearance parameters. This approach includes Active Shape Models ([Cootes et al., 1995](#)), Active Contours Models ([Kass et al., 1988](#)), Template Matching ([Yuille et al., 1992](#)), and Active Appearance Models ([Cootes et al., 2001](#)). Nevertheless, most of these methods are not enough robust against noise or variation on the image (pose, illumination, scale). On the other hand, the estimation of shape parameters is approached as a regression problem instead of model-fitting approach, where a cascade of regressors predicts the object shape or the location of its parts ([Burgos-Artizzu et al., 2013](#)). However, there are some difficulties which are not correctly managed by these methods. In the case of the model fitting method, they are not enough robust against noise or variation on the image (pose, illumination, scale). As regards regression-based models, it refines an initial guess of the landmark coordinates iteratively, thus the first initialization is critical. As a solution to above problem, a robust approach based on a cascade of regressors for facial point detection with three levels of convolutional networks (CNNs) was proposed by [Sun et al. \(2013\)](#). In this dissertation, the model-fitting method (Active Appearance Models) and the cascade of regressors based on CNNs have been tested for the face alignment.

In this dissertation, the face alignment was used for two task: face registration using deep

learning-based features in section 6.6, and face parts identification by means of the localization of facial points, also referred to as landmarks. Then, those facial landmarks were used to extract craniofacial measurements as described in sections 6.3, 6.4 and 6.5.

6.3 Local features

In this part, we present the uncalibrated craniofacial measurements used for OSA assessment. It includes the description of the procedure for measurements' extraction along with the analysis and results for OSA severity prediction. As described in the introduction of this chapter, these craniofacial measurements are based on those presented in (Lee et al., 2009a, 2010). Part of the research performed in this section was published in (Espinoza-Cuadros et al., 2015).

6.3.1 Facial landmark detection

Previous to the features extraction, it is important to identify a set of relevant facial landmarks on digital photographs. As shown in Table 3.1, the database of facial images contains frontal and profile digital photographs of 525 male subjects, that is 1050 images needed to be processed to obtain landmarks. Manual annotation as performed in (Lee et al., 2009a) can be tedious and it is prone to errors due to subjectivity. Consequently, we decided to use a widely used automatic landmarking method based on Active Appearance Model (Cootes et al., 2001). Based on the prior knowledge of landmarks position, the Active Appearance Models (AAMs) combines a statistical model, which represents the variation of shape and gray-level appearance of the object of interest, in this case, a face, with a gradient-descent fitting algorithm. Therefore, to train AAM models for frontal and profile photographs, we used a grid of 52 landmarks (frontal), taken from a general face identification system, and a 24 landmarks (profile), including specific marks for the neck area, respectively (Figure 6.2).

During the training stage of AMMs, frontal and profile appearance models were built from a set of manually annotated photographs using the **aam_tools**¹ software in order to learn the relationship between model parameter displacements (i.e. shape and texture) and the residual errors between a training image and a synthesized model example. To match to a test image, the AAM Search algorithm iteratively corrects the appearance parameters by minimizing the difference between a new image and one synthesized by the appearance model, leading to a better fit.

To get a more accurate landmarks annotation by the AAMs, previous to the landmarking

¹https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/software/am_tools_doc/index.html

Figure 6.2: Landmarks on frontal and profile view extracted using the *aam_tools*.

process, the photographs were converted to grayscale images due the non-controlled photograph capture process such as illumination conditions, the normalization of grayscale images' intensity was required, therefore, the process of adjusting intensity value was done automatically using histogram equalization. The process of feature facial landmarks annotation is depicted in Figure 6.3.

Figure 6.3: Block diagram describing the landmarking process by model fitting approach.

Although the AAMs performs well for representing shape and texture variations of an object, the model is local location-sensitive to face's position. In this Thesis, this effect is increased since photographs were not taken following a highly controlled procedure, such as illumination conditions, control of distance from the camera, and control of frontal and profile position. Accordingly, a human-supervised stage was also applied in order to correct some large deviation in the automatically generated landmarks. Once all facial landmarks were detected, the feature extraction was performed. Those features are described below.

6.3.2 Uncalibrated measurements

The main difference between (Lee et al., 2009a) and our research is that in the former study photographs were captured using a professional camera and following a highly controlled procedure. For instance, in the case of frontal position, it was achieved controlling that subject's facial landmark nasion was aligned along the subject alignment plane while ensuring both ears were seen equally from the front. As regards profile position, a laser pointer head-clip pointing to calibrated markings on the side wall was used to ensure the profile views were perpendicular to the frontal views. In addition, a calibration procedure to convert pixel measurements to metric dimensions was not included. As described in (Lee et al., 2009a), this was achieved by measuring some surface facial measurements, where some landmarks were previously identified by sticking a calibration marker in the patient's face (e.g. a circular nylon washer).

This highly controlled procedure contrast to our objective of exploring the possibility of using speech signals and images captured in a more usable scenario using standard portable devices, and looking for a better user experience avoiding manual identification (e.g. by palpation) of facial landmarks. Consequently, based on (Lee et al., 2009a) and (Lee et al., 2009b), and considering our limited photography capture process, three uncalibrated measurements were designed in order to capture the information associated with the craniofacial characteristic in OSA subjects. It is important to point out that in the case of relative distances between landmarks points Euclidean distance was used. As regards angles, degrees were used. These photographic measurements are detailed below.

6.3.2.1 Cervicomental contour area

General obesity is considered as one of the most important anatomical risk factors for OSA, in particular, the fat deposition on the anterior neck (Mortimore et al., 1998). In (Lee et al., 2009a) this anatomical risk factor along with soft tissue deposition is captured by four calibrated craniofacial photographic measurements: *Cervicomental angle*, where an increase of neck fat deposition causes an increase of this angle; *Cricomental space area*, *Thyromental space area*, and *Anterior neck space area*, where smaller areas values were found in OSA subjects due to space occupied by soft tissues or fat on the anterior neck.

Considering our limited photography capture process, it is extremely difficult to identify these points such as the cervical point or mentum, which was pre-identified on subjects by palpation. Accordingly, we have defined an alternative measurement, more robust to both our image capture and automatic landmarking process. This measurement was defined using a contour in the

cervicomenal region traced by six landmarks placed equidistantly (i.e. landmarks 11, 12, and 20-23 depicted in Figure 6.4) and annotated with high reliability following our semiautomatic AAM method. Therefore, the relative measurement of fat and soft tissue deposition on the anterior neck was calculated as the area of the cervicomenal region to the area of the rectangular region ratio. The area of the cervicomenal region consists of the gray solid line defined by landmarks 11, 12, and 20-23, and the bottom right vertex landmark V of the rectangle, as depicted in Figure 6.4 (left). The area of the rectangular region consists of the black dashed line defined by the bottom left landmark 23 and upper right landmark 11 ratio (Figure 6.4 (left)). This results in an uncalibrated measurement with a value that decreases as the fat deposition on the anterior neck increases.

Figure 6.4: Measurements used for the cervicomenal contour area (left) and measurements used for midface width (right).

6.3.2.2 Midface width

Lee et al. (2010) studied relationship between surface facial dimensions and upper airway structure in OSA subjects by means of the analysis of magnetic resonance images. Significant correlations were detected between surface facial dimensions and UA structures, in particular, *midface width* and *interocular width*. Lee et al. (2009a) reported midface-width-related measurements such as *face width-midface depth angle* and *face width-lower face depth angle*, where larger values were reported in OSA subjects. On the basis of these findings, we defined the *midface-width uncalibrated measurement* as the *midface width* to *interocular width* ratio. The corresponding landmarks and measurements are depicted in Figure 6.4 (right).

6.3.2.3 Tragion-ramus-stomion angle

Lowe et al. (1995) reported that OSA subjects had retracted mandibles, which is related to the inclination of the occlusal plane and the angle between the relative position of the maxilla to the mandible. Previous to Lowe et al. (1995), Jamieson et al. (1986) reported some craniomandibular abnormalities in OSA subjects such as retroposition of the mandible. In the same manner, the lowering of the hyoid bone often in conjunction with unusually small posterior airway space measurements were reported in OSA patients (Guilleminault et al., 1984). Those findings might suggest that several bony structures surrounding the oropharynx can be involved in the anatomic disarrangement leading to OSA.

On the basis of (Jamieson et al., 1986), (Guilleminault et al., 1984) and (Lowe et al., 1995), we proposed an uncalibrated measurement intended to capture, to some extent, the characteristic mandible position or mandibular retraction in OSA individuals. To define this angle (i.e. in degrees), we selected a set of landmarks that not only are related to the posterior displacement of the mandible but also could be accurately detected by our automatic landmarking process on the photographs without the need of prior marking. This measurement is the angle between the line ramus-stomion (see landmarks 16 and 6) and the ramus-tragion (see landmarks 16 and 18) as depicted in Figure 6.5 (left). Due to the displacement of the mandible by the retraction of the mandible, a slight displacement of the stomion, the landmark point 16 in Figure 6.5 (right), might be expected in OSA subjects, thus differences related to these angles between OSA subjects and controls are expected. As described before, the easy detection of involved landmarks makes easy their automatic detection, but it lacks precision as these craniofacial points are usually identified by palpation as performed in (Lee et al., 2009a).

Figure 6.5: *Tragion-ramus-stomion angle (left) and lateral cephalometric images of normal and OSA subjects (right). Right picture from (Jamieson et al., 1986).*

Once uncalibrated craniofacial photographic measurements were calculated, their relationships with OSA were analyzed following a statistical approach. Moreover, similar to speech features in Chapter 4, the evaluation of these measurements for OSA assessment was performed using them as input feature vector to a classification model. Therefore, the feature vector was building by concatenating the three uncalibrated measurements, as depicted in Figure 6.6.

Figure 6.6: Block diagram describing the OSA severity classification using uncalibrated craniofacial photographic measurements.

6.3.2.4 Analysis of uncalibrated measurements

Once all measurements were extracted, the statistical analysis was performed. Student's t-test was used for comparing OSA subjects and controls. Linear relationships between uncalibrated craniofacial measurements, clinical variables, and AHI were also examined by means of the Pearson correlation. The contrast analysis of uncalibrated craniofacial measurements between control and OSA subjects is presented in Table 6.1.

Table 6.1: Contrast analysis between Control and OSA groups for uncalibrated craniofacial measurements on the male population, including the mean and standard deviation ($\pm SD$).

Uncalibrated measurements	Control ($n = 160$, $AHI \leq 10$)	OSA ($n = 365$, $AHI \geq 10$)	p -value
Cervicomental contour area	0.63 ± 0.08	0.57 ± 0.09	< 0.001
Midface width	1.44 ± 0.06	1.46 ± 0.06	< 0.001
Tragion-ramus-stomion angle	114.42 ± 6.72	113.38 ± 6.21	0.09

As reported in Table 6.1, significant differences were found in the *cervicomental contour area* and *midface width*. In the case of *cervicomental contour area*, as described above, this uncalibrated measurement aims to describe the fat deposition and soft tissue on the anterior neck, and considering that overweight is an OSA risk factor, it could explain the relationship between this measurement and OSA. As regards the *midface-width*, greater values were found in OSA subjects. In the case of the *tragion-ramus-stomion angle* measurement, no significant differences were found between OSA and Control group. Besides the lack of precision on landmarks detection, these results could also be related to the proposed measurement, which does not represent the mandible's retroposition.

It is well-known the close relationship between these measurements and clinical variables, in particular, the *cervicomenta contour area*, which is related to the overweight and the high risk of suffering OSA. To evaluate this effect, a statistical contrast was performed for Control and OSA groups on a subset without differences neither on body mass index ($26 \leq \text{BMI} \leq 31$) nor on age, and age ($33 \leq \text{age} \leq 44$), as presented in Table 6.2.

Table 6.2: Contrast analysis between Control ($\text{AHI} < 10$) and OSA ($\text{AHI} \geq 10$) groups in the male population for uncalibrated craniofacial measurements on a subset without differences on body mass index and age.

Feature	Control ($n = 19$, $\text{AHI} < 10$)	OSA ($n = 24$, $\text{AHI} \geq 10$)	p -value
Clinical variables			
Age (years)	41.26 ± 1.59	40.66 ± 1.43	NS
Weight (kg)	90.15 ± 9.96	87.37 ± 8.52	NS
Cervical perimeter (cm)	41.26 ± 2.86	41.46 ± 1.63	NS
Body mass index (kg/m^2)	28.57 ± 1.48	28.21 ± 1.50	NS
Uncalibrated measurements			
Cervicomenta contour area	0.61 ± 0.09	0.58 ± 0.06	NS
Midface width	1.44 ± 0.07	1.44 ± 0.06	NS
Tragion-ramus-stomion angle	113.25 ± 5.81	112.73 ± 7.75	NS

Abbreviation: NS, non-significant.

As shown in Table 6.2, no significant differences between both groups were found. As regards *cervicomenta contour area*, the result was expected since once the effect of general obesity was removed no differences were found. However, according to (Lee et al., 2009a), the accumulation of fat or soft tissue on the anterior neck is not only constrained to general obesity. This fact was demonstrated in (Lee et al., 2009a) after BMI effect was removed and compared to the primary analysis of craniofacial photogrammetry data the measurement *anterior neck space area*, which represent the area of space in front of the neck and below the jaw, remained larger in OSA group. As regards *midface width*, no longer differences were reported between the control and OSA groups after matching for body mass index. In contrast, in (Lee et al., 2009a), the midface-related differences, in particular, the *Face width-midface depth angle* remained after BMI was removed.

In the same manner, linear relationships between uncalibrated craniofacial measurements, overweight-related variables (BMI, cervical perimeter), and Apnea-Hypopnea Index (AHI) were also examined in the entire OSA population. As reported in Table 6.3, the high correlations were found between the cervicomenta contour area and the clinical variables including the AHI. As regards relationship with the OSA severity, the weak correlation ($r = 0.36$) between

Table 6.3: *Pearson’s correlation between uncalibrated craniofacial measurement, clinical variables, and AHI on the male population.*

Uncalibrated measurement	AHI	BMI	Cervical Perimeter	Age
Cervicomenta contour area	-0.36	-0.55	-0.57	-0.18
Midface width	0.16	0.23	0.18	0.11
Tragion-ramus-stomion angle	-0.15	-	-	-0.14

Note: For clarity, non-significant correlation values were omitted. Correlation coefficients are significant beyond the 0.001 level of confidence.

the measurement and the AHI might be influenced by the modest relationship between that measurement and the BMI ($r = -0.55$, p -value < 0.001) and the modest correlation between the BMI and AHI ($r = 0.42$, p -value < 0.001), as shown in Table 5.1. In the same manner, the modest correlation between the *cervicomenta contour area* and cervical perimeter confirms the dimensional measurement of fat deposition on the anterior neck. In addition, the weak correlations between the uncalibrated measurements and age, it is linked to the increase of prevalence in elderly due to changes on upper airway structure such as increased deposition of fat in the pharyngeal area, lengthening of the soft palate, and changes in the body structures surrounding the pharynx (Albert et al., 2007).

After the effect of BMI was removed, similar to Table 6.2, the correlations between uncalibrated measurements and AHI were no longer significant. But the *midface width measurement* remained weak ($r = 0.33$, p -value $= 0.03$) and modest ($r = 0.40$, p -value $= 0.01$) correlations with cervical perimeter and BMI, respectively. These results make sense as the subset without differences on age and body mass index contains overweight male subjects ($BMI \geq 23$).

Therefore, the proposed uncalibrated measurements fail to represent the craniofacial characteristics of OSA population to some extent as compared to (Lee et al., 2009a). This suggests the lack of accuracy for representing facial characteristic might be due to semi-automatic landmarking process, which is clearly prone to errors. Nevertheless, the weak and modest correlations with OSA severity and clinical variables (i.e. BMI, cervical perimeter), respectively, support the notion that some craniofacial photographic measurements are capturing some facial soft tissue or adiposity, which are rather linked to general and regional obesity.

6.3.2.5 OSA severity classification

Similar to speech features in Chapter 4 and following the classification approach, the OSA assessment using craniofacial measurements were performed by means of the classification of subjects according to their OSA severity. During the test stage, each subject was first processed by extracting its uncalibrated craniofacial measurements and then features extracted were stacked

Table 6.4: OSA severity classification for male patients using clinical variables and uncalibrated craniofacial photographic measurements. Results from (Lee et al., 2009b) and (Balaei et al., 2017) are also presented.

Models	Acc (%)	Sen (%)	Spec (%)	AUC (95% CI)
Clinical variables (our study)	68.76	70.13	65.62	0.74 [0.69, 0.78]
Clinical variables (Lee et al., 2009b) ¹	76.10	86.00	59.10	0.78
Uncalibrated measurements (our study)	67.20	70.17	60.52	0.69 [0.64, 0.74]
Uncalibrated measurements (Balaei et al., 2017)	64.20	65.40	59.49	-
Uncalibrated measurements (Lee et al., 2009b)	71.10	80.70	54.50	0.80
Uncalibrated measurements +clinical variables (our study)	68.21	70.17	63.81	0.73 [0.67, 0.77]
Uncalibrated measurements +clinical variables (Lee et al., 2009b) ²	79.40	85.10	69.70	0.87
Calibrated measurements (Balaei et al., 2017)	63.70	67.6	55.0	0.69
Calibrated measurements (Lee et al., 2009b)	76.10	67.6	55.0	0.82

Note: Confidence intervals of AUC are not reported in (Lee et al., 2009b) and (Balaei et al., 2017).

¹ clinical variables={age, body mass index, witnessed apneas}.

² clinical variables={witnessed apneas, Mallampati class}.

on a fixed-length vector. Later, these vectors are used over the trained SVM model to predict the OSA severity of the test subject, as depicted in Figure 6.6. In order to compare the system performance and our craniofacial features, similar to (Lee et al., 2009b), we trained an SVM model using the variables clinical as input vector (age, body mass index, cervical perimeter, weight, height), where the most are good predictors of OSA, as performed in section 5.1 (see Table 5.3). The classification results are presented in Table 6.4.

Based on (Lee et al., 2009a) and (Lee et al., 2009b), a similar study was carried out by Balaei et al. (2017), where they replicated the same methodology (e.g. photographic capture process, manual landmarking, manual identification by palpation) along with the best craniofacial features (uncalibrated and calibrated measurements) for OSA prediction, as presented in (Lee et al., 2009b). However, the experiments were carried out on a different dataset of frontal and profile photographs of OSA diagnosed patients. Therefore, in Table 6.4, OSA prediction results by using uncalibrated measurements are shown along with Lee et al. (2009b) and Balaei et al. (2017) results. In addition, OSA prediction results from clinical variables and the fusion at feature level were also added.

According to results presented in Table 6.4, in terms of AUC, our approach reached a performance of 0.69 (95% CI [0.64, 0.74]) using uncalibrated measurements, 0.74 (95% CI [0.69, 0.78]) using only clinical variables, and 0.73 (95% CI [0.67, 0.77]) using the fusion at level feature of clinical variables and uncalibrated measurements. Similar performances for uncalibrated measurements are also reported by (Balaei et al., 2017). As described above, the photographic

capture and manual identification process followed in (Balaei et al., 2017) is the same as (Lee et al., 2009a), but results are different from those ones reported in Lee et al. (2009b). This finding makes relevant our results to some extent since we achieved similar performances to (Balaei et al., 2017) but in less controlled conditions, which is close to practical scenarios. Nevertheless, those findings disclose that results in Lee et al. (2009b) are not generalizable, as shown in Table 6.4, where lower performances were obtained in (Balaei et al., 2017) than (Lee et al., 2009b) using similar measurements.

Moreover, the loss of accuracy in our prediction model might be due to the less controlled photographic process and the landmarking process over the photographs. Besides that, it could be worth analyzing other reasons for this loss of accuracy. There may be some differences in the populations under study, in particular, the population explored by Lee et al. (2009b) includes both males (76.1%) and females (23.9%), while in our study male subjects were only considered. In addition, in (Lee et al., 2009b) the information of balance of female and male subjects in OSA and Control groups was not provided, and considering the significantly lower prevalence of OSA in women (Guilleminault et al., 1988) along with the craniofacial OSA risk factors differences between female and male (Chi et al., 2011), there is a high chance for the existence of machine-learning pitfalls in (Lee et al., 2009b). It is interesting to remark that in (Lee et al., 2009b), only two clinical variables were found to improve the classification accuracy when combining calibrated craniofacial measurements and clinical variables: *witnessed apneas*, that is the bed partner reporting witnessed apnea in the questionnaire, and *Mallampati classification*, that is a visual classification of the anatomy of the oropharyngeal structure used as screening method for OSA assessment. As pointed out in (Lee et al., 2009b), this results suggests that calibrated measurements are more important predictor than age and BMI (see Table 6.4), whereas witnessed apneas and Mallampati score provide complementary information to the classification model. Considering less controlled scenarios such as non-controlled photograph capture process, and the less invasive capture process for the patients to a great extent, to take advantage of uncalibrated measurement could be suitable for those conditions. Despite they are not good predictor such as clinical variables in our study, the information of witnessed apneas could be easily obtained from the patient itself in order to be used as complementary information to the model. In the case of the Mallampati score, its assessment by means of image processing techniques on subject's oropharyngeal region photography could be interesting in less controlled scenarios and less invasive for the patients.

In brief, despite the use of the uncalibrated measurements to face the uncontrolled photography capture condition, it seems that they are not quite discriminative to represent the different

OSA severities due to they are noteworthy influenced by the clinical variables, therefore they do not provide additional information. In this line, another relevant point to discuss, as already stated in (Lee et al., 2009b), is that despite measurements relate mainly to size rather than shapes, the latter may provide additional insights into the OSA craniofacial phenotype.

6.4 Statistical-model-based features

Besides the craniofacial measurements, another method for representing the craniofacial phenotype in OSA population is the shape and gray-level appearance modeling of objects by means of the Active Appearance Models (Cootes et al., 2001), nevertheless, we have focused on the shape variation as it can provide complementary information to uncalibrated measurements. The shape and gray-level appearance modeling along with the training and search AAM algorithm are described in detail in (Cootes et al., 2001).

6.4.1 Active Appearance Model (AAM)

Based on a priori knowledge of landmark points, the AAM combines a statistical model, which represents the variation of shape and gray-level appearance of the object with a gradient-descent fitting algorithm. In this Thesis, we only approached the shape variation as it could provide additional information to uncalibrated measurements. In addition, the gray-level appearance has also been tested, but no consistent results were obtained. Despite normalized photographs at the gray level, the poor performance of gray-level-appearance-based features might be due to the non-controlled photography capture such as illumination condition, which could contribute to the variability. Since the shape and the gray-level appearance modeling follow a similar training process, in this Thesis, only a brief description of statistical shape modeling is provided below.

Given a set of training of labeled images with their landmarks points previously annotated, we can generate a statistical model of shape variation, where the landmarks points distribution describe the shape of the object. Previous to the shape modeling, the landmarks are aligned into a common coordinate frame and represent each by a vector x . Let x_i be a vector describing the N points of the i th shape in the training set:

$$x_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, x_{i2}, y_{i2}, \dots, x_{ik}, y_{ik}, x_{iN-1}, y_{iN-1}) \quad (6.1)$$

The alignment process is performed by scaling, rotating and translating the training shapes to that they correspond as closely as possible. This process is described in detail in (Cootes et al., 1995). Each sample of the training set can be represented by a single point in a $2n$

dimensional space as presented in expression 6.1. Therefore, a set of S training samples gives S example shapes, which gives a cloud of S points in the $2n$ dimensional space. These points lie within some region of the space, space referred to as Allowable Shape Domain (Cootes et al., 1995). Accordingly, these points give an indication of the shape and size of this region, thus, new shapes can be generated in a systematic way according to new landmarks' positions within the Allowable Shape Domain. This approach can be summarized as they stated in (Cootes et al., 1995):

This approach attempts to model the shape of this cloud in a high dimensional space, and hence to capture the relationship between the positions of the individual landmark point. We make the assumption that the cloud is approximately ellipsoidal, and proceed to calculate its center (giving a mean shape) and its major axes, which give a way of moving around the cloud.

On the basis of Cootes et al. (1995), once all landmarks points were aligned to represent the shape variation of the ellipsoid, similar to the first step, the deviation from mean is estimated for each shape,

$$dx_i = x_i - \bar{x} \quad (6.2)$$

where dx_i is the deviation and $\bar{x}$ the mean. Then the $2N \times 2N$ covariance matrix W is calculated,

$$W = \frac{1}{N} \sum_{i=1}^N dx_i(dx_i)' \quad (6.3)$$

From the covariance matrix W , the principal modes of variation of the points of the shape are described by the eigenvectors (p_k , $k = 1, 2, \dots, 2n$), and they are estimated such that,

$$Wp_k = \lambda_k p_k \quad (6.4)$$

where p_k is the k th mode of variation of the shape and λ_k is the k th eigenvalue of W . The largest eigenvalues correspond to those eigenvectors of covariance matrix that describe the longest axes of the ellipsoid, thus the most significant modes of variation.

Once the principal modes of variation were calculated, any point x in the Allowable Shape Domain can be estimated through the mean $\bar{x}$, and adding a linear combination of the eigenvectors, as it is summarized below,

$$x = \bar{x} + P_s b_s \quad (6.5)$$

where $P = (p_1, p_2, \dots, p_t)$ is a set of orthogonal modes of variation and $b = (b_1, b_2, \dots, b_t)'$ is a vector of weights, also called set of shape parameters.

In the same manner, the gray-level and appearance modeling follows a similar procedure, which is described in detail in (Cootes et al., 2001). In the Active Appearance Models (AAMs) approach, the model is generated by combining a model of shape variation with a model of the gray-level appearance variations instead of using NOOOOOOOO only shape variation such as the Active Shape Models. The gray-level appearance model is summarized as

$$g = \bar{g} + P_g b_g \quad (6.6)$$

where $\bar{g}$ is the mean normalized gray-level vector, P_g is a set of orthogonal modes of variation and b_g is a set of gray-level parameters. Therefore, the shape and gray-level appearance of a sample can be summarized by the vectors b_s and b_g .

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} \quad (6.7)$$

where W_s is a diagonal matrix of weights for each shape parameter. To obtain the principal modes of variation, the PCA is applied on these vectors, giving a new model

$$b = Qc \quad (6.8)$$

where Q are the eigenvectors and c is a vector of appearance parameters controlling both shape and gray-levels of the model. Due to the linear nature of the model b , it can be expressed by the shape and gray-level model as a function of c as

$$\begin{aligned} x &= \bar{x} + P_s W_s Q_s c \\ g &= \bar{g} + P_g Q_g c \\ Q &= \begin{pmatrix} Q_s \\ Q_g \end{pmatrix} \end{aligned} \quad (6.9)$$

6.4.2 Face shape variation modeling

Based on the uncalibrated craniofacial measurements, described in the previous section, we have evaluated the variation of the shape of three facial regions related to the cervicomenta contour area in Figure 6.4 (left) and midface-width in Figure 6.4 (right). The descriptions of those facial

region shapes are provided below.

6.4.2.1 Cervicomental region

Similar to the uncalibrated measurement in Figure 6.4 (left), the cervicomental contour defined by landmarks points 11, 12, and 20-23, as depicted in Figure 6.7, aims to capture the variation of the cervicomental region.

Figure 6.7: *Cervicomental region.*

6.4.2.2 Midface-biocular region

Based on the uncalibrated measurement *Face width* in Figure 6.4 (right), this region is defined by the landmarks points 1, 11, 14, 18 drawing a trapezoid region, as depicted in Figure 6.8 (left). Since larger dimension in both midface and binocular width were found in OSA subjects, correlations between some shape variation modes and OSA are expected.

6.4.2.3 Midface region

According to (Lee et al., 2009a), some surface facial features such as midface width along with mandible-related dimensions (i.e. width and length) reported significant differences by contrasting OSA and Control groups. Therefore, we proposed a region defined by the landmarks points 0-12, as depicted in Figure 6.8 (right), to collect the variability related to those measurements, but adapted to our non-controlled photograph capture process.

Figure 6.8: *Midface-biocular region (left) and midface region (right).*

6.4.2.4 Analysis of shape-modeling-based features

According to (Cootes et al., 1995), the variance explained by each eigenvector of the covariance matrix is equal to the corresponding eigenvalue, therefore, the first t modes of variation explain most of the variance of shape. The contribution of each mode of variation related to the shape is determined by their weights. For instance, in Table 6.5 the percentage of variance of each parameter from the *midface-biocular region* is provided, where first shape parameters contain most of shape parameters' values to represent each mode of shape variation of the training dataset.

Table 6.5: *Percentage of deviation of shape parameters respect to the total deviation of shape parameters from midface-biocular region.*

Shape parameters	$\frac{\sigma(b_s)}{\sigma(b_r)} \times 100$
b_1	51.21%
b_2	20.37%
b_3	19.09%
b_4	9.33%

Note: σ =standard deviation as measurement of deviation of shape parameters.

Some shape parameters related to first modes showed some linear relationship with clinical variables and the Apnea-Hypopnea Index (AHI), as presented in Table 6.6, where the highest, though weak, correlation was between AHI and first shape parameter of *cervicomental region*. This parameter certainly represents the variation of area occupied in front of the neck and below the jaw, as shown in Figure 6.9. Similar to the *cervicomental contour area* measurement,

it captures the facial soft tissues or adiposity, and is also linked to general obesity, as shown by correlations with body mass index ($r = -0.54$, p -value < 0.001) and cervical perimeter ($r = -0.59$, p -value < 0.001). Furthermore, the very strong correlation between first shape parameter of *cervicomenal region* and *cervicomenal contour area* measurement ($r = 0.93$, p -value < 0.001) explains the similarity between them.

Table 6.6: *Pearson’s correlation between shape parameters, clinical variables, and AHI on the male population.*

Shape parameters	AHI	BMI	Cervical Perimeter	Age
Cervicomenal region (b_1)	-0.36	-0.54	-0.59	-0.24
Midface-biocular region (b_3)	-0.13	-0.21	-0.15	-0.10
Midface region (b_2)		0.16	0.16	

Note: For clarity, non-significant correlation values were omitted. Correlation coefficients are significant beyond the 0.001 level of confidence.

In the case of *midface-biocular region*, the third parameter shows a very weak correlation with AHI, but a strong correlation with *midface-width* measurement ($r = -0.81$, p -value $< .001$). As regards age, the weak correlation between shape parameters and age, the modes of shape variation of the *cervicomenal region* and *midface-biocular region* capture some of changes in upper airway structures due to aging such as fat deposition in the pharyngeal area, lengthening of the soft palate.

In addition, these correlations were no longer significant in a subset of matched subjects on body mass index (BMI) and age. Indeed, modest correlations reported between shape parameters, cervical perimeter and BMI were also no longer significant. In contrast, the relationship between uncalibrated measurements and those clinical variables remained significant, as described in section 6.3.2.4. Therefore, despite most of the subjects were overweight in the subset, these general obesity-related characteristics were not captured by those modes of shape parameters.

6.4.2.5 OSA severity classification

In order to evaluate the shape parameters to predict OSA severity in subjects, similar to the scheme in Figure 6.6, the shape parameters of the three facial regions were concatenated resulting in a fixed-length vector to be used as input to an SVM model.

From results in Table 6.7, the shape parameters ($AUC = 0.69$; 95% CI [0.60, 0.71]) achieved a similar performance to uncalibrated measurements ($AUC = 0.69$; 95% CI [0.64, 0.74]). In the same manner, the fusion of shape parameters and clinical variables did not provide relevant information to the model, as shown in Table 6.7. This result might be due to the existence of modest relationships between shape parameters and clinical variables (e.g. body mass index and

Figure 6.9: *Effects of varying the first shape parameter of the cervicomental region.*

cervical perimeter). Furthermore, considering the very strong correlation ($r > 0.9$) between the shape parameters and uncalibrated measurements, as described in previous section, the fusion of both did not provide additional information to the model.

From results above we could conclude that, besides the low performance of both types of features for OSA prediction, considering the less controlled photographic capture process and the human-supervised stage for landmarking annotation, it is reasonable to think that those factors could also have contributed to the low performance of both. In particular, the introduction of the human-supervised stage makes the landmarking detection based on model fitting more sensitive to errors resulting in a low performance of both types of features: shape variation parameters and uncalibrated measurements. Furthermore, as described in section 6.2.1, the landmark detection methods based on model-fitting (AAM) are not enough robust against noise or variation on the image (pose, illumination, scale) in consequence the less controlled photographic capture process makes this method prone to errors.

Table 6.7: *OSA severity classification for male patients by using shape parameters, and fusion at feature level with clinical variables and uncalibrated craniofacial photographic measurements.*

Feature	Acc(%)	Sen(%)	Spec(%)	AUC(95% CI)
Shape parameters	64.97	70.17	53.28	0.69 [0.60, 0.71]
Clinical variables	68.76	70.13	65.62	0.74 [0.69, 0.78]
Shape parameters + clinical variables	69.02	70.17	66.44	0.74 [0.69, 0.78]
Uncalibrated measurements + shape parameters	66.19	70.17	57.23	0.68 [0.62, 0.73]

6.5 Facial landmarks identification based on Deep Convolutional Network Cascade

Based on the results above, a more objective landmarking annotation process is required in order to avoid the errors due to the human intervention. Considering our less controlled pho-

tographic capture process and the limitation of the model-fitting-based methods are not robust to image variation (e.g. illumination conditions, and poses), a solution for those drawbacks is the face alignment by means of the regression approach based on convolutional networks cascade proposed in (Sun et al., 2013). This method overcomes some weakness of classical automatic landmarking approaches such as model-fitting (Cootes et al., 2001) and regression-model (Burgos-Artiz et al., 2013), in particular, the regression-model-based method where the update of landmarks positions by means of an iterative procedure depends on good initializations, which an incorrect initialization could lead to a local minimum. Moreover, many of visual extracted features are not reliable enough to estimate landmarks points due to the image variability presented on facial images (e.g. illumination conditions, pose variation, etc.). Therefore, the solution aims to use the information of the whole face region, in particular, the texture context information such that, at first level, the convolutional networks (CNNs) take the full face as input to extract global high-level of features at higher layers of the deep network based on the texture context information. Then, these deep features are used to predict effectively landmarks even when there are challenging conditions on images. The remaining two levels of the CNNs are used to refine the initial estimation.

In this dissertation, we have extracted facial landmarks from frontal face photographs by using a commercial face recognition system Face++, in particular, the Face++ Landmarks API², which is based on Convolutional Cascade Network cascade (Sun et al., 2014). For the experiments, we have used the official Matlab SDK for Face++ Landmarks, which locates and returns with high-precision landmarks of face components, including face contour, eye, eyebrow, lip and nose contour. All these facial regions are summarized by 68 facial landmarks, as depicted in Figure 6.10.

It is important to point out that before the facial landmarks detection, a face detector and aligning respect to horizontal axes was applied to get a high accuracy on landmarking process. For both process a MatlabTM toolbox for face recognition was used; the PhD toolbox³. In the case of aligning stage, the frontal facial photographs were aligned with respect to the horizontal axis. As reference landmarks for aligning pupils, the facial landmarks extracted by *aam_tools* were used as they can be accurately detected by the semiautomatic landmarking process on the photographs without the need of prior marking (landmark points 17 y 13 in Figure 6.2).

²<https://www.faceplusplus.com/landmarks/>

³http://luks.fe.uni-lj.si/sl/osebje/vitimir/face_tools/PhDface/

Figure 6.10: Landmarks on frontal view extracted by Face++ Landmarks API. Labelled landmarks (red points) correspond to midface width (points 4 and 13) and interocular width (points 22 and 72).

6.5.1 Performance evaluation

In order to compare both facial-landmark-annotation methods, and considering we have only tested automatic facial landmarking method on frontal facial photograph, we have extracted the uncalibrated measurement *midface width* (section 6.3.2.2) from landmarks extracted by both methods (semiautomatic method vs. automatic), and contrasted correlations with clinical variables and AHI, as presented in Table 6.8.

Table 6.8: Pearson's correlation between semi-automatically (*aam_tools* + human-supervised stage) and automatically (*Face++*) determined uncalibrated measurement midface width, clinical variables and AHI.

Method	AHI	BMI	Cervical perimeter	Age
aam_tools + human-supervised stage	0.16	0.23	0.18	0.11
Face++	0.18	0.29	0.29	0.19

Note: All correlation coefficients are significant beyond the .001 level of confidence.

From above table, high correlations values were obtained by uncalibrated measurements determined by the automatic method (*Face++*). In the same manner, we have also compared the predictive capability of two landmarking methods for OSA prediction. Therefore, we extracted the shape variation modes of the *midface region* (see section 6.4.2.2) and *midface-biocular region* (see section 6.4.2.3). Following a similar scheme for OSA prediction in Figure 6.6, the shape parameters were converted into fixed-length vectors and used as inputs to an SVM classifier and results are presented below:

As reported in Table 6.9, it shows a high accuracy of automatic facial landmarks identi-

Table 6.9: Classification results for OSA prediction by semi-automatic method (i.e. *aam_tools*) and automatic method (*Face++*) and using shape parameters from midface and midface-biocular regions. Results from (Balaei et al., 2017) for OSA prediction by using an automatic landmarking method are also shown.

Method	Acc(%)	Sen(%)	Spec(%)	AUC(95% CI)
aam_tools + human-supervision	60.07	70.25	37.34	0.53 [0.47, 0.59]
Face++	63.20	70.25	47.46	0.66 [0.61, 0.71]
Uncal. measurements (Balaei et al., 2017)	67.30	67.50	67.20	-
Cal. measurements (Balaei et al., 2017)	69.80	68.50	76.40	0.69

Abbreviation: Cal, calibrated; Uncal, uncalibrated.

fication (*Face++*) as compared to semi-automatic method (*aam_tools*), where a noteworthy improvement, in terms of AUC, indicates that features are able to extract OSA-related facial information. Furthermore, similar results were obtained in (Balaei et al., 2017). These results indicate the lack of generalization of the prediction model in (Lee et al., 2009b), in particular, the craniofacial measurements due to the manual annotation of landmarks, despite the high-controlled procedure of annotation, it is still prone to errors.

Besides that, a question arises: it is possible to predict OSA by using both a machine learning model, without the intermediate step of measuring the craniofacial features, and the facial images as input to this model?

6.6 Deep learning-based features

In recent years, the continuous development in different computer vision areas has allowed dealing with challenging tasks in image recognition; those tasks related to real scenarios rather than controlled scenarios, where most of the databases were designed to facilitate the study of specific parameters of the image recognition. In the context of face recognition, those controlled conditions (e.g. position pose, lighting, expression, background, etc.) have allowed developing reliable and accurate local image descriptors in order to extract discriminative information from image; local descriptors such as Scale-invariant feature transform (SIFT), Local-binary patterns (LBP), and Histogram of oriented gradients (HOG). Nevertheless, more and more most of challenging tasks in image recognition aims for non-controlled scenarios, that is real scenarios where images were captured without controlled conditions such as lighting, position, pose, etc. Nowadays, most of the image databases exhibits natural variability in pose, lighting, facial expression, background, and resolution, along with a large number of images. For instance, for object recognition, there is The ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015) dataset, which is a benchmark in object category classification and detection on hundreds

of object categories and millions of images. In the case of face recognition, there are WDRF (Chen et al., 2012), CelebFaces (Sun et al., 2014), and VGGFace (Parkhi et al., 2015) databases. In the same manner, new deep learning-based approaches have been developed, which, in most of the cases, outperformed the handcrafted features (SIFT, LBP, etc.). In the case of face recognition, one of the most used deep architectures is the Convolutional Neural Networks (CNNs) and a representative system is the DeepFace method (Taigman et al., 2014).

Besides the successful performance of the deep learning for facial recognition, in recent years other successful applications have achieved similar performances, for instance skin cancer diagnosis (Esteva et al., 2017), which could reduce the cost of diagnosis as well as the time-consuming analysis, or in the social psychology field such as the analysis of facial patterns of people according to the sexual orientation (Wang and Kosinski, 2017). Although this study can bring out some controversy, from a research scope, it can provide information about how the deep networks “see” or extract the information from an image dataset.

In the context of this dissertation, the exploration of deep learning-based features for OSA prediction have been motivated by following points:

- The non-controlled photography capture process: besides the image pre-processing stage was applied on photographs (normalized, aligned), the landmark detection based on model fitting are not enough robust against noise or variation on the image (pose, illumination, scale) in consequence the local facial descriptors (uncalibrated measurements, shape parameters) are no reliable.
- The use of handcrafted features for face recognition: the intermediate step to extract features, which relies on the accuracy of landmarks, could be prone to errors. In this Thesis, besides both types of features, inspired by (Lee et al., 2009a), aim to extract facial patterns associated with OSA phenotype, from results in previous sections, the “pre-designed” craniofacial measurements did not show generality.
- The deep architectures along with the large quantity of training data from labeled face databases have allowed outperforming the benchmark systems based on handcrafted features. In this dissertation, we were not able to reach that requirement due to the small number of training samples to properly train a deep network. Similar cases have been faced in the state-of-the-art computer vision. As a solution to this issue, many deep learning-based methods for feature extraction have been studied such as the well-known *Transfer Learning*. It is important to point out that the concept of Transfer Learning does not only focus on the deep learning-related problems or to solve the problem related to the small

number of training data, but it also embraces more applications related to classification, regression, and clustering problems (Pan and Yang, 2010).

6.6.1 Transfer Learning

The use of deep architectures for image recognition, as described above, involves a high quantity of training data such as thousands of labeled faces for training the CNNs. For instance, the ImageNet ILSVRC model was trained on 1.2 million images (Krizhevsky et al., 2012). In the case of object recognition, considering the scenario where we want to classify any object such as classifying model of cars, lung disease from X-ray images, facial expressions instead of the common classification task. In addition, considering that the quantity of available data in those specific classification tasks is usually lower as compared to object recognition, it results in an impossible task.

As a solution to this issue, many studies have tackled those limitations under the idea of extracting the knowledge from one or more source tasks and applies the knowledge to a target task. That idea is referred to as Transfer Learning Pan and Yang (2010). Although this concept involves an extended definition as well as their different application on machine learning problems, which are out of the scope of this dissertation, thus we only have approached the definition involved for deep learning-related problems, in particular, those related to object recognition task.

Therefore, in the context of the deep learning, the underlying idea of the Transfer Learning consist in two steps: first, pretrain a deep network (net) on a very large dataset. Second, use the pretrain net either as an initialization or a fixed feature extractor for the task of interest. There are two typical forms or applications for Transfer Learning:

- **pre-trained net as fixed feature extractor:** on this form, the last layers from the net are removed, then the rest of the net is used as a fixed-length feature extractor for the new dataset. Once features were extracted from all the dataset, a linear classifier (e.g. SVM, SVR) can be trained for the new dataset. One of the most relevant works in computer vision tasks by using this approach is the research performed in (Razavian et al., 2014), where authors tested this form of Transfer Learning in many visual recognition tasks such as object image classification, scene recognition, fine-grained recognition, attribute detection and image retrieval, outperforming most of the benchmarks of those visual recognition tasks.
- **Fine-tuning:** similar to the above approach, the top last layers from the net are replaced

and retrained on the new dataset, but the weights of the pre-trained network are also fine-tuned by continuing the backpropagation. There are many forms of fine-tuning weights, for instance, fine-tuning all the layers of the pre-trained net or fine-tuning some higher-level portion of the net. This approach has been also tested in (Razavian et al., 2014), achieving similar performances to benchmarks.

In this dissertation, on the basis of (Wang and Kosinski, 2017) and (Razavian et al., 2014), we have explored the Transfer Learning on pre-trained-net-as-feature-extractor mode to represent the craniofacial phenotype of OSA subjects.

6.6.2 VGG-Face-CNN as feature extractor

From the Transfer Learning concept, the goal is to obtain a generic image representation extracted from a pre-trained deep net in order to apply those generic features to different recognition tasks. Although many studies have confirmed this concept applied on small number of visual recognition tasks, Razavian et al. (2014) carried out a study on a wide range of recognition tasks. The study confirmed the generic descriptors extracted from a CNNs can be used as a generic image representation to approach different recognition tasks. These extracted features are used for training a linear classifier such as SVM for the specific classification task. On the basis of the principle of using the CNNs as feature extractor, Razavian et al. (2014) tested different configurations by using the output of each layer from a pre-trained CNNs as input features to a classifier (e.g. linear SVM) to perform different recognition tasks and different datasets. Moreover, they also performed fine-tuning that is replacing and retraining the classifier of the top of the net, and fine-tune the weights of the pre-trained network via backpropagation. Both approaches achieved benchmark or near benchmark performances on different vision tasks such as object recognition, scene classification, attribute detection, scene image retrieval, and others. In the case of image classification, the findings in (Razavian et al., 2014) suggested that information at earlier layers of the net contains more generic features such as edge detectors or color blob detector that should be useful to many tasks, but later layers becomes progressively more specific to the classes contained in the dataset.

6.6.2.1 Problem formulation

In the context of this dissertation, considering our small number of samples in dataset, the use of a pre-trained net as a feature extractors fits well for our scenario, where we aim to extract generic information from the earlier layers from the net, and use them as input features to

train a linear classification model (i.e. Support Vector Machine) to OSA severity classification. For this purpose, we used the pre-trained network VGG-Face CNN (Parkhi et al., 2015). This architecture consist of 16 layers (13 convolutional layers and 3 fully-connected layers). The network was trained by using large-scale dataset: 2.6 million images from 2622 people. The input to the network is a face image of size 224x224.

Therefore, as depicted in Figure 6.11, the information was extracted from earlier layers of the net, in particular the first-fully-connected layer (fc6), that is the layer 14 from the VGG-Face CNN. It is important to point out that this the first fully-connected layer achieved the best performance during the experiments. Certainly, this aspect is expected due to the last fully-connected layers of the net contains specific related information to the training data instead of generic characteristics. This aspect was also reported in (Razavian et al., 2014) and further demonstrated in 6.11. Once all features were extracted, they are used as input to the classifier (i.e. SVM). From now on these types of features are referred to as deep features.

As depicted in Figure 6.11, the size of the input image to the net is 224×224 . Therefore, our frontal photographs were resized to required dimension. In order to get a more accurately extraction, all photographs were aligned. Thus, the faces were centered and rotated to align the eyes horizontally according to their landmarks' location, previously extracted by using the Face++.

Figure 6.11: Scheme of VGG-Face CNN descriptor used as feature extractor. The features are extracted from the full-connected layer 6. Then, these features are employed as input to the classifier so as to predict OSA in subjects.

6.6.3 Deep features validation

Previous to the OSA assessment, following a similar scheme in Figure 6.11, the deep features were employed to predict some clinical variables (age, weight, BMI, etc.) to validate those features extracted from the VGG-Face. Therefore, the problem formulation aims to predict clinical variables' values by using the deep features and a prediction model (Support Vector Regression, SVR).

Table 6.10: *Clinical variables estimation for male patients using the deep features extracted from VGG-Face.*

Clinical variable	MAE	Correlation Coefficient
Age (years)	4.76	0.87
Cervical perimeter (cm.)	2.11	0.57
Body mass index (Kg/m ²)	3.26	0.52
Weight (Kg.)	11.47	0.50
Height (cm.)	5.63	0.14

Note: All correlation coefficients are significant beyond the 0.001 level of confidence.

Results in Table 6.10 indicate that deep features are good descriptors to represent physical characteristics in faces such as age ($r = 0.87$, p -value < 0.001), cervical perimeter ($r = 0.57$, p -value < 0.001) and body mass index ($r = 0.52$, p -value < 0.001). Similar to these physical characteristics collected by these descriptors, the characterization of OSA facial phenotype is expected. Since these descriptors capture facial soft tissue or adiposity as well as aging, which are associated with OSA risk factors, thus, it is possible the extraction of OSA facial characteristics.

6.6.4 OSA assessment by deep features

Following the scheme depicted in Figure 6.11, we evaluated the extracted features by means of the transfer learning for OSA severity classification in subjects. For this experiment, we followed the classification approach described in section 3.3.1.2 to compare the final performance with other facial representation methods, in particular, those based on the handcrafted features such as the uncalibrated measurements. A Support Vector Machine classifier (SVM) was employed following the classification approach as described in section 3.3.1.2. In addition, we have also compared performance of later layers of VGG-Face, in particular the fully-connected layers 6 and 7.

As reported in Table 6.11, a high performance was achieved by the first fully-connected layer (layer 14). Certainly, this confirms that deeper layers contain specific information related to the training data instead of the generic information such as the intermediate layers. This

Table 6.11: *OSA severity classification results for male patients by using deep features extracted from later layers of VGG-Face.*

Deep features	Acc(%)	Sen(%)	Spec(%)	AUC(95% CI)
Fully-connected6 (layer 14 in VGG Face)	66.86	70.11	59.49	0.68 [0.62, 0.73]
Fully-connected7 (layer 15 in VGG Face)	65.50	70.11	55.06	0.65 [0.60, 0.71]

result indicates that the deep features are able to represent some characteristics from the OSA population similar to uncalibrated measurements. In addition, both methods achieved similar performances in terms of AUC, as depicted in Figure 6.12. Nevertheless, the reached performance could have been influenced by clinical variables since they are good predictors of those variables, as shown in 6.10, hence the influence of clinical variables (age, cervical perimeter, etc) on the deep features.

Figure 6.12: *ROC for OSA detection by using uncalibrated measurements and deep features.*

From above results, we fulfill the first two motivations explained in section 6.6, where the problems related to face landmarks detection, and the extraction of handcrafted features are no longer a matter of concern by using the Transfer Learning techniques. In other words, by means of this approach, an end-to-end scheme can be performed using only the image as input to the net, where generic, but relevant and “unbiased” information can be extracted from the earlier layers of the network.

6.6.4.1 Face surface regions employed by the classifier

Following the methodology proposed in (Wang and Kosinski, 2017), we analyzed which part of the facial images enabled the OSA prediction. Therefore, we masked parts of the facial image and measuring the degree to which the prediction performance has changed. If a given area of the facial image is important to the model, a significant change in the performance is expected (and vice-versa).

Using an area of 28×28 pixels in the top-left corner was masked in all images. The procedure was repeated 64 times while sliding the mask across the grid covering the entire image, composed of 8×8 squares (each sized at 28×28 pixels). The change in performance resulting from masking a given area of the image is measured by the absolute difference between the mean absolute error (MAE) of the masked and original image. The results are presented as heat maps in Figure 6.13 showing the degree of performance change due to the masking. Moreover, this procedure was extended to the clinical variables in particular age and weight, which are known as good predictors of OSA as they are related to OSA risk factors as aging and general obesity.

According to the visual analysis from Figure 6.13, heat maps confirm that the prediction relied on the facial area whereas the background does not provide relevant information for AHI and clinical variables prediction.

As regards weight estimation, the relevant areas include the cheeks, forehead and infraorbital surface, and the most informative area, according to heat maps intensity, the volume of fat deposition on the anterior neck, which clearly played a role, along with the cheeks in less extent, to determine if a person suffers obesity. This condition also affects the forehead and infraorbital surface. In the same manner, as regards cervical perimeter, the relevant areas are infraorbital and forehead surface in less extent. Nevertheless, the relevant area is the central region of the nose. This aspect was also reported by Lee et al. (2010), where they found a relationship between the cervical perimeter and the nose width.

As regards age prediction, the informative facial areas include the nose, mouth, and lips. According to (Albert et al., 2007), most of the changes by aging are related to skeletal and soft-tissue facial contour changes due to the anterior facial skeleton trend to greater convexity with age, thus the soft tissues of the face become more protuberant. Among these changes there are the growth in the nose and ears, lengthening of the lips, and mandibular length increase.

Finally, as regards AHI estimation, among the informative facial areas, forehead surface, midface region, intercanthal distance, lower-face region, and mandibular region showed the most color intense value of heat maps. Some of these facial parts were also documented in (Lee et

Figure 6.13: Heat maps showing the degree to which masking a given part of an image changes the performance for the AHI and clinical prediction in the male population. The color scale ranges from the blue (no change) to red (substantial change). The color-coded squares were smoothed using 2D Gaussian filtering.

al., 2010), specifically midface width and lower-face width reached significant correlations OSA. Nevertheless, similar to speech features, the relationship between OSA and these informative facial areas might be influenced by clinical variables, hence the relationship between those facial regions and OSA.

6.6.4.2 Facial features analysis in OSA prediction

In this section, we have analyzed the differences in facial features between Control and OSA faces that enabled the classification in the previous section (6.6.4). Therefore, we compared the faces classified as Control or OSA. This analysis consists in the analysis of composite faces and the average landmarks' location.

Previous to the analysis, a process of face alignment was performed. The faces were centered and rotated to align the eyes horizontally according to their landmarks' location, previously

extracted by using the Face++. From aligned faces, new landmarks were extracted and averaged for each group; those classified as Control ($AHI < 10$) or OSA ($AHI \geq 10$). In the same manner, from aligned and centered faces, the values of corresponding pixels were averaged across images so as to produce composite faces.

As depicted in Figure 6.14, by visual inspection some differentiating facial characteristics can be found between non-OSA and OSA subjects, in particular, those related to overweight, aging, and tiredness due to the sleepiness in OSA subjects. Therefore, these findings suggest that there are some facial patterns associated with the OSA severity in subjects.

Figure 6.14: *Composite men faces built by averaging faces classified as OSA and Control by using the deep features.*

In the same manner, differences in craniofacial parts were found between Control and OSA groups, as shown by average landmarks of frontal photographs in Figure 6.15. For instance, contour drawn by average landmarks location (5-8 and 14-17) showed that OSA patients have wide jaws. This finding is supported by significant correlations achieved between the relative distances between landmarks (relative distance defined as the distance in pixels between landmarks coordinates 5-14, 6-15, 7-16, 8-17 normalized by the interocular width in landmarks coordinates 22-72), as presented in Table 6.12. Such distances also show significant correlations with the body mass index (BMI) and the cervical perimeter. Again, those significant correlations suggest the influence of clinical variables on those craniofacial and agree with some discoveries in (Lee et al., 2009a), where they reported significant moderate correlations between mandible width and OSA ($r > 0.4$).

Table 6.12: *Correlation values between landmarks' distances from mandibular region, BMI, Cervical perimeter and the AHI.*

Relative distances between landmarks	AHI	BMI	Cervical perimeter
5-14	0.24	0.31	0.31
6-15	0.22	0.30	0.31
7-16	0.19	0.25	0.28
8-17	0.15	0.19	0.23

Note: All correlation coefficients are significant beyond the 0.001 level of confidence.

6.7 Summary

The results of OSA assessment using craniofacial descriptors indicate that these features are able to carry some information related to specific characteristics in OSA faces. Although the influence of clinical variables on those features could explain the relationship with OSA, that behavior is expected to a great extent since the OSA predisposition is highly determined by the risk factors such as obesity, aging, gender and others. Furthermore, it is important to point out that although our results are below to those reported in "reference" studies (Lee et al., 2009a,b), the study of Balaei et al. (2017) revealed that results in cited studies are not generalizable.

Figure 6.15: *Average facial landmarks on frontal view extracted from male population's facial photographs and using Face++ Landmarks API. Labeled landmarks correspond to mandibular region (i.e. points 5-8, 14-17) and interocular width (points 22 and 72).*

In the case of the local features, Lee et al. (2009a,b) showed that it can be useful to extract some surface facial measurements from the facial photographs. In the same context, but in non-controlled photography capture conditions (head position, distance from the camera, lightning, etc.), we were able to extract information related to OSA faces by using uncalibrated craniofacial measurements and features based on shape-appearance modeling. However, the results indicated their performances are very sensitive to the process of extraction of the features, in particular,

the identification of landmarks on the faces. In our case, we followed a semi-automatic process of landmarks annotation based on the Appearance Active Models (AAMs). Since this method is location-sensitive to face's position the introduction of the human-correction stage was necessary, which is prone to errors due to the subjectivity. Using a completely automatic landmarks annotation process and using the shape variation parameters from the *midface-biocular region*, the model performance improved respect to the initial one, approximately 20%. A similar approach was also followed by (Balaei et al., 2017), where they reached similar results to ours by using uncalibrated and calibrated craniofacial features but following a controlled photography capture process and automatic facial landmarks annotation process instead of manual.

As regards deep features, similar performances to uncalibrated craniofacial measurements were achieved. Contrasting both types of features, by means of the deep learning features we avoid the errors introduced by the facial landmarks identification and the use of the “pre-designed” craniofacial measurements based on (Lee et al., 2009a), which did not show generality, as shown in (Balaei et al., 2017). Nevertheless, the most relevant aspect is that we were able to perform a “quasi” end-to-end model for OSA prediction based on the deep learning.

Chapter 7

Obstructive Sleep Apnea Syndrome in Women

IN THIS CHAPTER we present the study of speech and craniofacial characterization techniques for OSA prediction in women. The study comprises the analysis of the speech and craniofacial features used in the male population (Chapters 4 and 6), and the evaluation of both types of features following the same experimental framework (see section 3.3).

7.1 Introduction

The prevalence of OSA in women is lower than men: 3% for women and 10% for men between 30-49 years old and 17% for men and 9% for women between 50-70 years old (Peppard et al., 2013). The way of how OSA syndrome affects to female population differs in the symptoms, diagnosis, and consequences. Many studies documented gender differences in the upper airway structure, fat deposition distribution and respiratory stability in OSA. Besides the OSA risk factors reported in men population such as aging and central obesity, there is the hormonal factor, which contributes to the growth of prevalence, but it decreases as age increase (Wimms et al., 2016).

As regards OSA symptoms, women reported more generalized daytime symptoms than men; symptoms such as insomnia, restless legs, depression, nightmares, palpitations, and hallucinations. In contrast, men are more likely to report snoring and apneic episodes (Valipour et al., 2007). Most of typical OSA symptoms such as snoring are not usually reported by female patients during the diagnosis since they may consider it irrelevant or the information from a partner on snoring and witnessed apnea is not provided as they attend clinical appointments

on their own (Wimms et al., 2016). The frequent reporting of uncommon OSA symptoms in women (e.g. headache, anxiety, and depression) along with the low prevalence might lead to the under-evaluation of OSA the female population (Young et al., 1996). In a community-based study (Young et al., 1996), women diagnosed with OSA reported similar symptoms across a range of severities, being the snoring the best predictor of OSA. Nevertheless, in (Young and Finn, 1998), they found greater gender-related differences in OSA symptoms expression at lower AHI values, in particular, symptoms such as snoring, witnessed apneas, and excessive daytime sleepiness.

Epidemiological studies documented gender-related differences in the upper airway (UA) structure, fat distribution, and respiratory stability. In the case of the UA's structure, by means of MRI imaging, differences between gender were found. For instance, the airway length, the soft palate, the tongue, and the total amount of soft tissue in the throat are smaller in women than in men (Anttalainen et al., 2013). In the same manner, the probability of UA collapsibility is higher in men due to their own configuration, that is longer and softer oropharynx, and fatter and larger more posterior tongue, increasing the susceptibility of the UA to collapse (Lin et al., 2008).

In the case of central obesity, considered a risk factor for OSA, in OSA population women tend to be more obese than men at the same level of AHI (Jordan et al., 2005). This effect is related to the difference in fat distribution between both genders. For instance, men tend to have greater total neck soft tissue volume and more fat deposition around the neck than women (Whittle et al., 1999). Those differences were also found in our dataset, as shown in Table 7.1, where women show lower overall AHI, weight, and cervical perimeter values than men, in contrast, body mass index (BMI) shows a slightly higher value in women as compared to men.

Table 7.1: Summary on the women and men, including the mean and standard deviation (mean $\pm$ SD) for each of the clinical variables and AHI. The last column includes p -values estimated by non-parametric Wilcoxon rank-sum test applied to each variable considering only women and men.

Variable	Women (n=204)	Men (n=525)	p -value
AHI	15.11 $\pm$ 17.52	22.63 $\pm$ 18.43	< 0.001
Age (years)	51.42 $\pm$ 11.65	49.01 $\pm$ 12.34	0.01
Cervical perimeter (cm.)	36.85 $\pm$ 2.93	42.00 $\pm$ 3.17	< 0.001
Body mass index (Kg/m ²)	30.37 $\pm$ 6.86	29.60 $\pm$ 5.02	0.09
Weight (Kg.)	78.72 $\pm$ 17.50	90.95 $\pm$ 16.95	< 0.001
Height (cm.)	161.16 $\pm$ 6.36	175.17 $\pm$ 7.16	< 0.001

The hormonal factor also plays a role in the prevalence of OSA in women. According to (Dancey et al., 2001; Young et al., 2003), differences in prevalence between both genders decrease

by aging due to a marked increase in the prevalence and severity of sleep disorders breathing after menopause in women. Furthermore, Valipour (2012) suggests that progesterone levels may play a role in protecting women from sleep apnea in premenopausal since it provides a protective effect on upper airway patency since it is a respiratory stimulant and increase the upper airway muscle tone as well as they also play a role in the distribution of body fat, hence there is low OSA prevalence in premenopausal women but increase in prevalence after menopause (Block et al., 1980). Besides the known OSA risk factors, hormonal factors contribute to the way in how OSA affects each gender, in consequence, it must be considered as a confounding factor as well as differences between premenopausal and postmenopausal women.

7.2 Clinical variables as OSA predictors

The central obesity and aging are known as important OSA risk factors. In fact, a higher body mass index (BMI) and age indicate a greater severity of obstructive sleep apnea. Thus, the clinical variables related to those risk factor are good predictors for OSA. Nevertheless, since there are differences in how these factors affect each genders, most of these clinical variables show different relationships with OSA, as shown in Table 7.2.

Table 7.2: *Pearson's correlation between clinical variables and AHI on the female population.*

	AHI	Height	Age	Weight	BMI	CP
AHI	1					
Height		1				
Age	0.21	-0.36	1			
Weight	0.22	0.14		1		
BMI	0.27	-0.21		0.93	1	
CP	0.34	-0.16	0.20	0.63	0.68	1

Abbreviation: BMI, body mass index; CP, cervical perimeter.

Note: For clarity, non-significant correlation values were omitted. Correlation coefficients are significant beyond the 0.001 level of confidence.

From correlations in Table 7.2, correlations between AHI and clinical variables show lower values as compared to men in Table 5.1. These fact may anticipate the less influence of the clinical variables for our OSA female population. As regards age, it shows a slightly higher correlation with the AHI than men. A possible explanation of this effect, although it has to be exhaustively studied, might be the hormonal factor in women as it contributes to the OSA prevalence, in particular, the hormone progesterone is a respiratory stimulant which provides a protective effect on upper airway patency and/or ventilator drive (Valipour, 2012). This protective effect decreases as age increases. It is important to point out that we did not regard

the menopausal status for this Thesis since most of the women from our database are middle adult age (i.e. mean age 51.42 in Table 7.1), thus we virtually considered all women living beyond the age of 55 years experience menopause in consequence there are no significant differences in menopausal status in our female population.

Contrasting all correlations obtained for both populations, in terms strength correlation index, correlations in men are modest ($0.4 \leq r \leq 0.69$) while in women correlations are weak ($0.2 \leq r \leq 0.39$). In the case of men, the influence of overweight-related variables such as weight, body mass index, and cervical perimeter is high in men, in contrast, those clinical variables show a less influence in the female population, resulting in a degradation of performance for AHI prediction and OSA severity classification as presented in Tables 7.3 and 7.4 as compared to men in Tables 5.2 and 7.3.

Table 7.3: *Female subjects' AHI estimation by using clinical variables.*

Method	MAE	Correlation Coefficient
SVR linear kernel	10.11	0.30
SVR RBF kernel	10.52	0.27

Note: Correlation coefficients are significant beyond the 0.01 level of confidence.

Table 7.4: *OSA classification severity for female subjects by using clinical variables.*

Method	Acc(%)	Sen(%)	Spec(%)	DOR(95% CI)	AUC(95% CI)
SVM linear kernel	68.13	70.83	65.74	4.66 [2.57, 8.43]	0.72 [0.64, 0.79]
SVM RBF kernel	66.17	70.83	62.03	3.96 [2.20, 7.13]	0.69 [0.61, 0.76]

Contrasting performances in both approaches (regression, classification) between women and men (Tables 5.2 and 5.3), low performances for OSA assessment was achieved in the female population. The Bland-Altman plot in Figure 7.1 also provides information about how the regression model performs. In both genders, it can be seen that those patients with higher AHI values than 30 events/hr (severe OSA subjects) were underestimated, where there is a trend toward a plateau. Also, the Bland-Altman plot shows that the estimated AHI corresponds more closely to the true one for mean AHI values below 30 events/hr. These results might indicate a good performance for OSA screening, whereas the underestimation of severe patients, as long as they are diagnosed as severe it is tolerable. Nevertheless, the agreement between the estimated and the true AHI values by means the diagnostic agreement approach indicates differences between female and male population (Table 7.5). In the case of women, there is a higher percentage of agreement ($Ag = 70.59\%$) than male ($Ag = 54.67\%$). With the exception

of underestimated rates, better values of overestimated and no-agreement were obtained for women than men.

Table 7.5: *Diagnostic agreement approach for female and male subjects' AHI estimation by using clinical variables.*

Gender	Agreement	Overestimation	Underestimation	No-agreement
Female	70.59%	4.90%	18.14%	6.37%
Male	54.67%	14.10%	16.38%	14.85%

Those results suggests, in the case of overestimated rate, better accuracy for screening OSA at lower AHI values in women than men. In the case of no-agreement percentage, it might indicate how the model performs to some extent. Nevertheless, it should be noted that the diagnostic agreement approach does not assess the prediction performance but the screening ability for OSA compared to the reference model from a clinical perspective.

Beyond the knowledge of clinical variables performance, it would be interesting to see how each variable contributes to the prediction model. Although it is known that best predictors for OSA are age and BMI ([Hoffstein and Szalai, 1993](#)), the gender-differences also have an impact on best predictor for OSA. Therefore, we have employed a straightforward method, that is the forward-backward stepwise regression method ([Draper and Smith, 1998](#)). It consists in adding and removing variables from a linear model based on their statistical significance in explaining the response variable. The method begins with an initial model, at each step the method searches for the variable to remove or add from the model based on a criterion. We have employed the p -value of an F-statistic to test models with and without the variable under analysis at each step. To compare the gender-effect on most important clinical variables, we also have replicated the experiment for the male population. For both experiments, we employed the function `stepAIC` from the `StepAIC` package of R. As criteria for removing or adding a term, we have employed a p -value of 0.1 and 0.05, respectively, with a

Table 7.6: *Multiple linear regression model for AHI prediction in women. The model is fitted to the training data by means stepwise method.*

Linear regression model				
$y \sim 1 + AGE + CP$				
Estimated coefficients				
	Estimate	SE	tStat	p-value
Intercept	-0.74	0.42	-1.74	0.08
AGE	0.03	0.01	5.82	< 0.001
CP	0.04	0.01	4.05	< 0.001

Table 7.7: *Multiple linear regression model for AHI prediction in men. The model is fitted to the training data by means stepwise method.*

Linear regression model				
$y \sim 1 + BMI + AGE + CP + AGE \times CP$				
Estimated coefficients				
	Estimate	SE	tStat	p-value
Intercept	-10.31	2.08	-4.93	< 0.001
AGE	0.19	0.04	4.68	< 0.001
CP	0.26	0.05	5.01	< 0.001
BMI	0.04	0.01	4.08	< 0.001
AGE $\times$ CP	-0.00	0.00	-4.29	< 0.001

constant model as the starting model.

As reported in Tables 7.6 and 7.7, the set of clinical features selected are different for each gender. In the case of men (Table 7.7), those variables are age, BMI, and cervical perimeter, while in female (Table 7.6) the selected clinical features are only age and cervical perimeter. In consequence, the BMI has a higher impact on male than female population. It explains the differences in fat distribution between genders. The association between the BMI and OSA are more pronounced in men than women due to that physiological and mechanical effects have a positive impact on the upper airway in female population (Hoffstein and Szalai, 1993).

7.3 OSA assessment by speech features

Abnormality speech is expected due to anatomical and functional changes in the upper airway structure. Since there are differences in OSA-related clinical factors between genders, differences in how speech features perform for OSA prediction are also expected.

Most of the previous studies for OSA prediction by means of speech features have only focused on male subjects. To the best of our knowledge, there are no similar studies that concentrated on female OSA patients, and very few publications are available that discuss this issue (Goldshstein et al., 2011; Solé-Casals et al., 2014).

Therefore, we aimed to extend the exploration of speech features for OSA voice characterization in women. From results in Chapter 4, we used the most discriminative speech features that are the i-vectors.

7.3.1 Apnea-Hypopnea Index prediction

In this section, we present the results of AHI and clinical variables (Table 7.8). In the case of clinical variables with the exception of height variable, good performances were obtained, in

Figure 7.1: Regression and Bland-Altman plots for female a) and male b) speakers' AHI estimation by using clinical variables.

particular, for age and cervical perimeter. In the same manner, those variables also reached good performances in the male population (see Table 4.7).

As regards AHI prediction in Table 7.8, best performance was reached by low dimensionality i-vectors (dim 200; MAE = 10.48; $r = 0.28$; $p\text{-value} < 0.001$; 95% CI [0.15, 0.40]). As compared to the male population (dim=100; MAE = 13.63; $r = 0.30$; $p\text{-value} < 0.001$; 95% CI [0.21, 0.37]), there is no differences in performances in terms of correlations, but in terms of MAE, best result was obtained for women. This low MAE value might be influenced by the low overall AHI values in the female population. This aspect can be better visualized through the Bland-Altman and regression plots in Figure 7.2. Similar to clinical features, in both genders, patients with higher AHI values than 30 events/hr are underestimated, where there is a trend toward a plateau. In the case of male population, this trend is noteworthy since there are more severe cases ($AHI > 30$) than female population. This trend is also shown in the Bland-Altman plots, where the mean AHI values are away from the mean difference, which indicates a poor agreement with the

Table 7.8: Female speakers' AHI and clinical variables estimation by using *i*-vectors and SVR-linear kernel.

Variable	Mean Absolute Error (MAE)						Correlation coefficient					
	i-vector dimension						i-vector dimension					
	400	300	200	100	50	30	400	300	200	100	50	30
AHI	10.55	10.52	10.48	11.09	11.00	10.91	0.26	0.26	0.28	0.18	0.23	0.23
Height (cm.)	4.85	4.91	4.99	4.96	4.88	5.02	0.31	0.25	0.21	0.29	0.30	0.21
Age (years)	8.28	8.02	7.98	8.48	8.63	8.48	0.49	0.50	0.50	0.44	0.39	0.42
Weight (Kg.)	12.73	12.69	12.68	12.65	13.47	13.45	0.33	0.31	0.37	0.38	0.21	0.21
BMI (Kg/m ²)	5.02	5.14	5.06	5.06	5.28	5.40	0.35	0.29	0.34	0.32	0.23	0.21
CP (cm.)	2.09	2.07	2.06	2.01	1.97	2.13	0.43	0.47	0.45	0.48	0.45	0.36

Abbreviation: BMI, body mass index (AHI); CP, cervical perimeter.

Note: All correlation coefficients are significant beyond the 0.001 level of confidence.

Table 7.9: Diagnostic agreement approach for female and male speakers' AHI estimation by using *i*-vectors of dimension 200 for women and 100 for men, and SVR-linear kernel

Gender	Agreement	Overestimation	Underestimation	No-agreement
Female	69.12%	3.92%	20.59%	6.37%
Male	52.95%	13.52%	16.57%	16.96%

true AHI values. The agreement approach confirmed this lack of agreement between methods (Table 7.9), where a higher agreement rate is obtained in women (Ag = 69.12%) than men (Ag = 52.95%). Moreover, low overestimation and no-agreement rates of AHI are also showed in women might suggest the best accuracy for OSA screening at low AHI values.

7.3.2 OSA severity classification

Similar to experiments in chapter 4, we evaluated the performance of speech features based on the classification approach.

Table 7.10: OSA severity classification for female speakers by using *i*-vectors and SVM model (linear kernel).

i-vector dimension	Acc(%)	Sen(%)	Spec(%)	DOR (95% CI)	AUC (95% CI)
400	63.23	70.83	56.48	3.15 [1.76, 5.63]	0.70 [0.62, 0.77]
300	61.76	70.83	53.70	2.81 [1.57, 5.03]	0.68 [0.61, 0.75]
200	68.62	70.83	66.66	4.85 [2.67, 8.80]	0.73 [0.65, 0.79]
100	58.33	70.83	47.22	2.17 [1.21, 3.88]	0.66 [0.58, 0.73]
50	62.25	70.83	54.62	2.92 [1.63, 5.22]	0.69 [0.61, 0.75]
30	58.82	70.83	48.14	2.25 [1.26, 4.02]	0.67 [0.60, 0.74]

As reported in Table 7.10, similar to the regression approach, best result was obtained by using low dimensionality *i*-vectors (Acc = 68.62%, Sen = 70.83%, Spec = 66.66%, AUC = 0.73;

Figure 7.2: a) Regression and Bland-Altman plots for female speakers' AHI estimation by using *i*-vectors. b) Regression and Bland-Altman plots for male speakers' AHI estimation by using *i*-vectors. Points inside the rectangular area (black line) are the percentage of diagnostic agreement between the two methods.

95% CI [0.65, 0.73]), and a slightly better performance was reached as compared to the male population (Table 4.18) as represented in Figure 7.3 in terms of AUC.

Few studies have been performed for OSA detection by means of speech features in women such as (Goldshtein et al., 2011; Solé-Casals et al., 2014; Tyan et al., 2017). Goldshtein et al. (2011) performed the classification of OSA subjects by gender by using a short- and long-term features, extracted from vowels and nasal phonemes recordings. Best performance was reported by the short-term features (Sen = 74%, Spec = 71%). In (Solé-Casals et al., 2014), they carried out a gender-independent experiment, where the OSA and control groups contain very different percentages of male and female speakers. They achieved high classification rates (Sen = 81.49%, Spec = 84.69%). Despite good performances reported by both studies, there are some machine learning pitfalls in the performed methodology which could explain the good performance obtained such as the imbalance between women and men percentages which act as a

Figure 7.3: ROC for OSA detection in men and women by using *i*-vectors.

confounding factor that could explain the good classification results, as well as the use of feature selection techniques from a large number of acoustic parameters on small number of cases. In our recent publication (Tyan et al., 2017), we studied the relationship between frequency formants and bandwidths, extracted from sustained vowels, and the AHI. The most relevant finding showed significant correlations between acoustic features and AHI; analysis performed on a subset of the population without differences in age and BMI.

Table 7.11: AHI value prediction in a female and male population without differences in clinical variables by using *i*-vectors.

Prediction model	MAE	Pearson correlation	<i>p</i> -value
i-vectors200-women	11.13	0.14	0.26
i-vectors100-men	12.05	0.22	0.06

From above results, it suggests that part of performance in women might be influenced by the correlation between OSA and clinical variables, in particular, the age. To assess this influence in OSA prediction, a subset of female patients was filtered without differences in clinical variables. As shown in Table 7.11, results confirmed this influence of clinical variables on the speech features. Besides the degradation of AHI prediction performance, the significant correlation between predicted and true values might suggest that *i*-vectors represent information related to OSA without clinical variables' influence. Nevertheless, considering the small number of samples in the female population, this hypothesis could not be confirmed.

7.4 Craniofacial phenotype characterization in OSA women

As described in Chapter 6, craniofacial abnormalities have been reported in OSA men such as longer soft palate, dental occlusion, longer distance between the hyoid bone and the mandibular plane, large tongue base, relaxed pharyngeal soft tissues (Guilleminault et al., 1984), which increase the probability that UA collapses. Similar physical alterations have been also documented in female population (Guilleminault et al., 1988; Lin et al., 2008) such as differences between the length of the soft palate, distance from the mandibular plane to the hyoid bone. In addition, some studies have reported craniofacial differences within OSA women due to hormonal factor (Guilleminault et al., 1988), such as longer mandibular plane-hyoid distance in premenopausal than postmenopausal women; contrary to what would have been expected with aging (Kollias, 1999; Kollias and Krogstad, 1999).

In any case, since most of the women from our database are middle adult age (i.e. mean age 51.42 in Table 7.1), we have virtually considered all women living beyond the age of 55 years will experience menopause, consequently, no significant differences in menopausal status in our female population were expected.

Therefore, following the experimental framework performed in the male population (Chapter 6), in this part we have evaluated two types of craniofacial features: uncalibrated measurements and deep learning-based features.

7.4.1 Uncalibrated measurements

We have analyzed the relationship between these measurements, clinical variables, and AHI. Furthermore, contrast analysis of uncalibrated craniofacial measurements between control and OSA subjects was also carried out.

Table 7.12: Contrast analysis between Control ($AHI < 10$) and OSA ($AHI \geq 10$) groups for uncalibrated craniofacial measurements on the female population, including the mean and standard deviation ($\pm SD$).

Uncalibrated measurement	Control ($n = 108$)	OSA ($n = 96$)	p -value
Cervicomental contour area	0.61 ± 0.10	0.57 ± 0.09	0.007
Midface width	1.40 ± 0.04	1.42 ± 0.06	0.024
Tragion-ramus-stomion angle	112.86 ± 6.30	111.98 ± 6.25	0.32

As reported in Table 7.12, similar to the the male population (see Table 6.1), there are significant differences in the *cervicomental contour area* and *midface width* between Control and OSA groups. From correlations in Table 7.13 and contrasting both genders, the *cervicomental*

contour area measurement shows a better correlation in men ($r = -0.36$; p -value < 0.001 ; see Table 6.3) than women. According to (Whittle et al., 1999), the total neck soft tissue volume and the pharyngeal fat is higher in men than women, which explains the high correlation between that *cervicomental contour area* measurement and AHI in men. In contrast, *midface-width* shows a lower correlation in men ($r = 0.16$; p -value < 0.001 ; see Table 6.3). Actually, these differences between correlations might be associated with OSA women are more obese than men at the same AHI (Leech et al., 1988), consequently, the more fat around the midface, the larger the width.

Table 7.13: *Pearson's correlation between uncalibrated craniofacial measurement, clinical variables, and AHI on the female population.*

Uncalibrated measurement	AHI	BMI	Cervical perimeter	Age
Cervicomental contour area	-0.25	-0.61	-0.56	
Midface width	0.26	0.27	0.27	0.22
Tragion-ramus-stomion angle				-0.24

Note: For clarity, non-significant correlation values were omitted. Correlation coefficients are significant beyond the 0.001 level of confidence.

As regards *tragion-ramus-stomion angle* measurement, there are no significant correlation with AHI was found as well as no significant differences between OSA and Control group. Nevertheless, in both genders, this measurement and age show a weak and significant correlation. This finding is linked to the increase of prevalence in elderly due to changes on upper airway structure such as increased deposition of fat in the pharyngeal area, lengthening of the soft palate, and changes in the body structures surrounding the pharynx (Albert et al., 2007).

From above results, correlations indicate the relationship between these measurements and AHI might be influenced by clinical variables, which show significant correlations with AHI (Table 7.2). To evaluate this effect, a contrast analysis was performed for Control and OSA groups on a subset without differences neither on body mass index ($25 \leq \text{BMI}$) nor on age ($41 \leq \text{age} \leq 55$). As reported in Table 7.14 no significant differences are shown. Although there are no reference studies on female population, these results contrast with those in Lee et al. (2009a), where some significant differences between control and OSA groups remained after clinical variables effects were removed.

Similar to the experiments performed on the male population by using uncalibrated craniofacial measurements, following the same procedure described in Figure 6.6, we have evaluated these measurements for OSA prediction in women.

Results in Table 7.15 show lower performance for OSA severity classification in female population. Contrasting both genders, the low influence of clinical variables on the uncalibrated

Table 7.14: *Contrast analysis between Control ($AHI < 10$) and OSA ($AHI \geq 10$) groups in female population for uncalibrated craniofacial measurements on a subset without differences on body mass index and age.*

Feature	Control ($n = 37$)	OSA ($n = 27$)	p -value
Clinical variables			
Age (years)	49.13 ± 4.07	49.85 ± 4.43	NS
Weight (kg)	84.21 ± 19.18	88.14 ± 16.67	NS
Cervical perimeter (cm)	36.94 ± 1.97	38.11 ± 2.59	NS
Body mass index (kg/m^2)	32.50 ± 7.14	34.30 ± 6.65	NS
Uncalibrated measurements			
Cervicomental contour area	0.58 ± 0.09	0.54 ± 0.10	NS
Midface width	1.41 ± 0.04	1.42 ± 0.05	NS
Tragion-ramus-stomion angle	112.02 ± 5.85	113.51 ± 5.23	NS

Abbreviation: NS, non-significant.

Table 7.15: *OSA severity classification for boths genders by using uncalibrated craniofacial photographic measurements.*

Method	Acc(%)	Sen(%)	Spec(%)	AUC(95% CI)
Uncalibrated measurements (women)	52.97	70.52	37.38	0.54 [0.45, 0.61]
Uncalibrated measurements (men)	67.20	70.17	60.52	0.69 [0.64, 0.74]

measurements' performance of women could be the cause of the poor performance. Furthermore, since these measurements rely on the accuracy of the facial landmarking method, which is clearly prone to errors due to the human-supervision stage, consequently, this factor is likely to be the main source of low performance in uncalibrated measurements. Although this handicap can be overcome by employing an automatic facial landmarking method (section 6.5), these results suggest that those uncalibrated measurements are not suitable for craniofacial characteristic in women, even more if there are differences in craniofacial abnormalities between genders, in particular, the body fat distribution (Guilleminault et al., 1988) as well as the distribution of soft tissue and volume of fat on the neck (Whittle et al., 1999). In addition, the uncontrolled conditions of photography capture (e.g. distance to the camera, pose, illumination, etc.) make the measurements no robust against these conditions.

7.4.2 Deep learning-based features

From results obtained by the uncalibrated measurements, it suggests that, first, they are prone to errors due to subjectivity introduced by the human-supervised stage in the facial landmarking procedure. Second, the non-controlled photographic capture process contributes to the low

performance of those features.

Following the same procedure for extracting deep features described in Figure 6.12, in this section we present the results for OSA assessment by using the deep features extracted from the VGG-Face net (Parkhi et al., 2015).

7.4.2.1 Features validation

In order to validate the deep features extracted, we predicted the clinical variables (age, weight, BMI, etc.) by using these features. Since the VGG-Face is a face classifier, the information related to those clinical variables are expected to be included in the deep features.

Table 7.16: *Clinical variables estimation for female patients by using the deep features extracted from the pre-trained net VGG-Face.*

Clinical variable	MAE	Correlation Coefficient
Age (years)	5.43	0.82
Cervical perimeter (cm.)	1.98	0.58
Body mass index (Kg/m ²)	4.63	0.55
Weight (Kg.)	11.97	0.53
Height (cm.)	5.23	NS

As reported in Table 7.16, the deep features are good predictor of most of clinical variables with an exception of height. Similar to the male population, the best estimations were obtained for age ($r = 0.82$, p -value < 0.001), cervical perimeter ($r = 0.58$, p -value < 0.001), and body mass index ($r = 0.55$, p -value < 0.001).

Therefore, these features capture information related to facial soft tissue changes, adiposity, and aging, which are related to OSA risk factors, thus, it is expected that deep features were able to extract some facial information related to facial OSA phenotype.

7.4.2.2 OSA assessment by deep features

Following the scheme depicted in Figure 6.11, we have evaluated the deep features for OSA severity classification.

Table 7.17: *OSA severity classification for female and male patients by using deep features extracted from the full-connected₆ (Layer 14) of the VGG-Face CNNs.*

Model	Acc(%)	Sen(%)	Spec(%)	AUC(95% CI)
Fully-connected ₆ (women)	67.32	70.52	64.68	0.68 [0.60, 0.75]
Fully-connected ₆ (men)	66.86	70.11	59.49	0.68 [0.62, 0.73]

As shown in Table 7.17, similar performances for both genders were achieved. Nevertheless,

contrasting performances between deep features and craniofacial measurements best performance was obtained by the deep features, as depicted in Figure 7.4. This result stands out the remarkable performance of deep features over the local features, which are more prone to errors due to the landmarks detection, or, not suitable for a different population.

Figure 7.4: ROC for OSA detection by using uncalibrated measurements and deep features.

7.4.2.3 Analysis of facial surface elements for OSA assessment

Similar to analysis performed in section 6.6.4.1, we have examined which part of the facial surface enabled the OSA prediction. In order to address this question, we masked parts of the facial image and measuring the degree to which the prediction performance has changed. If a given area of the facial image is important to the model, a significant changed in the performance is expected (and vice-versa).

By means of visual inspection in Figure 7.5, heat maps confirm that the prediction relied on the facial are whereas the background does not provide relevant information for AHI and clinical variables prediction.

As regards the AHI estimation, the most informative facial areas include some parts of the midface, nose, and the anterior neck. In contrast, in the male population, the most informative area was the mandibular region (Figure 6.14).

As regards the weight estimation, the most informative facial areas included the nose, forehead and intercanthal-distance-related region. According to (Lee et al., 2010), those facial areas show significant correlation with the weight, for instance, the correlation between the nose-width

($r = 0.15$, $p\text{-value} < 0.05$), the intercanthal-distance ($r = 0.27$, $p\text{-value} < 0.05$) and the body mass index.

In the case of age, according to (Albert et al., 2007) among normal facial soft tissue changes by aging in the 50s, there are fine facial lines appearing horizontally across the forehead and vertical lines emerging between the eyebrows, which might be associated with the most informative area for the age prediction.

Finally, as regards cervical perimeter, the most informative region include the intercanthal-distance-related region, which according to (Lee et al., 2010), they show a significant correlation with the cervical perimeter.

Figure 7.5: Heat maps showing the degree to which masking a given part of an image changes the performance for the AHI and clinical prediction in the female population. The color scale ranges from the blue (no change) to red (substantial change). The color-coded squares were smoothed by using 2D Gaussian filtering.

Similar to the facial features analysis performed in section 6.6.4.2, from the composite faces in

Figure 7.6 we can observe some differentiating facial characteristics between non-OSA and OSA subjects, in particular, overweight, aging, and tiredness due to the sleepiness in OSA subjects.

Figure 7.6: *Composite men faces built by averaging faces classified as OSA and Control by using the deep features.*

7.5 Summary

There are differences in how OSA affects women, besides the typical risk factor reported such as aging, overweight, there is the hormonal factor that plays a role in the syndrome's development and also provides a protective mechanism against the upper airways collapse. Certainly, there are craniofacial and physical differences between OSA female and male population, which will determine the disease development; physical differences such as the fat body distribution or craniofacial differences such as longer soft palate, dental occlusion, longer distance between the hyoid bone and the mandibular plane, large tongue base, relaxed pharyngeal soft tissues, which increase the probability that UA collapses. Due to changes in the structure of the UA, speech abnormalities are expected.

As regards speech characterization in OSA subjects, i-vectors reached similar performance to the male population. As regards AHI prediction, in terms of MAE, a slightly higher result was obtained for women, which might be due to the low overall AHI values in our female population than men as higher AHI values prediction was prone to errors in both genders. For OSA severity classification, the best performance was achieved in the female population. In this case, the indirect influence of age on i-vectors stood out, being a 5% of improvement respect to performance in the male population. Nevertheless, once the effect of the clinical features was removed, there was a considerable decrease in performance.

Furthermore, there are differences in craniofacial features for OSA prediction between women and men. In the case of the uncalibrated craniofacial measurements, lower classification rates were reported in women than men. This finding clearly reveals that those uncalibrated features are not suitable for collecting craniofacial changes in OSA female patients due to craniofacial differences between genders. It is important to point out that those features are prone to errors due to the facial landmarks detection method followed, which certainly contributes to the low performance achieved by these features. In contrast, deep features reached higher performance than uncalibrated features, approximately 25% of improvement respect the latter ones as well as a similar performance to the male population. Again, the advantages provided by this type of features are noteworthy, besides removing the tiresome process of extracting the uncalibrated features along with the high chance of error due to the human-supervised stage, deep learning-based features have shown to be more robust against to uncontrolled conditions of photographic capture process, which makes it suitable for our scenario.

Chapter 8

Conclusions, Research Contributions and Future Work

8.1 Summary of the Thesis

THE DEVELOPMENT OF THE THESIS'S RESEARCH was motivated by different starting points scenarios, as described in Chapter 1. Each scenario is summarized below:

- **Scenario 1:** on the basis of the success of automatic speech processing methods and acoustic analysis of speech employed for OSA assessment, we aimed to explore the most successful and widely used for speech characterization in the state-of-the-art automatic speaker recognition, which is based on the utterances' spectral modeling: i-vectors and supervectors. To the best of our knowledge, it is the first time that these speech descriptors are employed for speech-based OSA detection. Those speech representations are based on utterance spectral modeling by means of the adaptation of a Universal Model GMMs (UBM) to a specific utterance spectral space, accordingly, the UBM have to explain the high variability captured from a population in order to reach the best adaptation, hence a great amount of training data is required for its training. In the context of this Thesis, we employed, in terms of hours, more than 20 hours of Spanish speech recordings from asymptomatic speakers and ~ 1.5 hours of symptomatic speakers (OSA subjects), which add some OSA-related variability to the acoustic space. The development of this scenario is described in Chapter 4.
- **Scenario 2:** most of the previous studies for speech-based OSA detection were performed over small datasets, and using high dimensionality features along with feature selection

algorithms. Good performances for OSA severity classification and AHI prediction were achieved by these studies. Some studies performed OSA detection via speech features on two types of populations: OSA symptomatic and asymptomatic, which could lead to false discoveries since most of the symptoms in OSA population such as snoring can be clearly distinguished between both populations through the speech. Furthermore, some studies performed gender-independent experiments, where female and male individuals were included in the same classification or prediction tasks. Therefore, we considered limitations and deficiencies when developing and validating machine learning techniques, leading to the overfitting and false discoveries. Regarding these facts, we performed a review of speech processing techniques for OSA detection employed in those studies. The development of this scenario is described in Chapter 5.

- **Scenario 3:** on the basis of the novel quantitative photographic approach for OSA assessment by means of a set of craniofacial measurements (Lee et al., 2009a,b), and the relationship between some abnormalities in the upper airway's structures and some surface facial parts (Lee et al., 2010) in OSA subjects, we aimed at the craniofacial characterization of OSA subjects by means of face characterization techniques inspired by Lee et al. (2009a) work such as uncalibrated measurements (angles, areas) and statistical-model-based features for OSA assessment. Furthermore, the breakthroughs in computer vision for tasks related to object representation (image classification, scene recognition, image retrieval) by means of deep learning-based techniques reached high performances, even some benchmarks have been overcome by such techniques. Considering our less controlled photography capture process, we aimed at the exploration of those craniofacial characterization and deep learning-based techniques for object characterization, in particular face characterization. The development of this scenario is described in Chapter 6.
- **Scenario 4:** most of previous studies for OSA assessment by means speech had focused on male subjects. Barring the the works of Goldshtein et al. (2011) and Solé-Casals et al. (2014) for OSA assessment, there are no similar studies that concentrated exclusively on women. A possible explanation of this fact may be related to the low prevalence of OSA in women, hence a limited number of subject to properly perform a thorough analysis. Nevertheless, the high percentage of undiagnosed cases also contribute to this low prevalence. Besides the typically OSA-related symptoms, there are other factors associated with OSA that only affect women such as the hormonal status. As regards craniofacial features, only Lee et al. (2009a) considered the differences between genders, but they

assumed that confounding due to gender differences was unlikely. Based on these facts, we have explored the speech and craniofacial characterization techniques for OSA assessment in the female population following a similar methodology performed in Chapter 4. The development of this scenario is described in Chapter 7.

8.2 Conclusions

From those starting scenarios, the research carried out in this thesis concludes:

- **Scenario 1:** lower performances were achieved by supervectors and i-vectors than previous studies by using similar features. Contrasting both types of features, i-vectors achieved the highest performance for OSA assessment. This result is associated with the i-vector approach to project the main direction of variability from high dimensionality Total Variability subspace into low dimensionality subspace; i-vectors. Best performance was reached by low dimensionality i-vector. This could be related to less speaker and channel variability in the utterances' recordings since all were collected by a controlled and similar recording conditions as well as utterances are constrained to a specific set of phrases, thus main sources of variability might be due to assigned to low dimensionality. Nevertheless, it is known that i-vector approach requires a high amount of training samples in order to properly train the total variability subspace. Despite our development dataset included only approximately ~ 1.5 hours of OSA-symptomatic patients, it was not sufficient to properly represent the main direction of acoustic variability in OSA patients. In addition, according to (Dehak et al., 2011a), in speaker verification, the i-vector approach comes together with a session variability compensation in order to compensate the session variability in the low dimensionality space rather than the high one, which is the GMM supervector space. In this context, on the basis of (Bahari et al., 2014) we carried out OSA compensation variability, however, no reliable results were obtained due to two possible reasons: 1) there are not enough samples for the WCCN matrix training, or 2) there is no discriminant acoustic information associated with OSA severity patients. Moreover, clinical features, in particular, those associated with OSA risk factors such as age, and BMI, which are good predictors of AHI. Nevertheless, the i-vectors and supervectors showed good performances for clinical variables estimation, in particular, age, cervical perimeter, and BMI. Consequently, part of performances reached by i-vectors and supervectors are influenced by those clinical features. In addition, two different approaches for training total variability subspace were addressed: 1) using as development dataset a large amount of short

Spanish utterances from asymptomatic and OSA speakers, and 2) using as development dataset only OSA speakers. Contrasting both approaches, results indicated slight degradation of performance in 2) respect to 1). Nevertheless, this aspect needs to be discussed further as more training samples are required in order to correctly train a total variability subspace. At this point, it is clear that we are not dealing with a voice pathology. Our findings suggest that, actually, there are no specific cues in OSA speech. Despite there are some specific acoustic characteristics that could be found in some patients for instance nasalization or hypernasalization, there are no generalized features found in voices affected by abnormalities in the upper airway structure due to the OSA.

- **Scenario 2:** results support the hypothesis of the presence of machine learning pitfalls into the validation process of machine learning techniques as well as limitations and deficiencies due to the wrong approach of these issues, such as the performance of feature selection algorithm over high dimensionality features and a small training set. Moreover, clear differences between OSA symptomatic and asymptomatic subjects reveals differences through the speech by using some acoustic features. Likewise, differences between genders can be also collected by speech features, which also revealed some false discoveries in some previous studies.
- **Scenario 3:** two different approaches for facial OSA patients' characterization were performed. Results indicated that uncalibrated craniofacial measurements captured some facial soft tissues or adiposity, which are in turn closely linked to general and regional obesity (e.g. cervicomentalar contour area) and aging to less extent. Likewise, same relationships were found by the statistical model based features, in particular, the shape-variation modeling of some facial regions. The main drawback of these local-features is that the performance is dependent on the facial-landmarking detection's accuracy. Despite there was a high improvement when contrasting automatic and manual facial landmarking approach, the improvement is still below the uncalibrated measurements in (Lee et al., 2009b). The deep-learning based approach, in particular, the transfer learning method shows a performance similar to local features, nevertheless, their advantages rely on extracting pixel information on the entire facial area instead of local regions previously defined by landmarks. In addition, this approach resulted more suitable to our scenario as there is more robust against some variability introduced by the uncontrolled photography capture process (i.e. pose, illumination, distance). Besides both approaches reached similar performances, the deep-learning-based features open a new front for OSA assessment

by facial descriptors, which it has demonstrated to collect some OSA patients-related facial characteristics without the necessity of the facial landmarking procedure.

- **Scenario 4:** there are clear differences in how OSA affects women and men. These differences are also manifested on upper airway structures configuration and craniofacial differences found in both populations. Moreover, the hormonal factor plays an important role in the OSA pathogenesis in women, which also contribute to these differences. An evidence of this effect might be linked to aging, where a higher correlation was found between age and the AHI in women than men. In the case of speech features, a similar performance to the male population was reached, but in terms of MAE, a slight improvement was reported for AHI prediction in the female population due to the lower overall AHI values compared to men, where higher AHI values (severe OSA) were under-predicted in both populations. In addition, the higher correlation with age in women than men might have influenced in the best OSA severity classification performance in the female population. In the case of craniofacial features, uncalibrated craniofacial measurements shown lower performance for OSA prediction than men. Similar to the male population, the semi-automatic facial landmarking process has an influence on the final performance of these measurements. The deep-learning based features reached a higher performance than uncalibrated features but contrasting these features between genders, similar performance is reached, nevertheless, different facial surface regions were considered by the final classifier for both genders, which supports the gender differences in the OSA pathogenesis.

8.3 Research Contributions

Following the analysis approach for each scenario, the contributions are the following:

- **Scenario 1 and 2:** the study carried out in both scenarios has contributed to both to emphasize the possible existence of potential pitfalls when using machine learning for medical diagnosis, and to guide further research on the connection between speech and OSA. In this latter aspect, we believe there is an open way to future research looking for new insights in this connection using different acoustic features, languages, speaking styles, or recording positions. However, besides properly addressing the methodological issues when using machine learning, any new advance should carefully explore and report on any possible influence of speech and AHI mediated through other clinical variables or any other factor that could affect speech as speakers' dialect, gender, or mood state.

- **Scenario 3:** the study of craniofacial features in OSA-diagnosed patients extracted from photographs under uncontrolled photographic capture conditions has opened a new front for a noninvasive preliminary assessment of OSA based on easily- and readable- collected data by employing powerful and robust against-image-variability technique based on the deep learning. In addition, the design of uncalibrated craniofacial measurements based on Lee et al. (2009a) research demonstrated the effectiveness to represent some surface facial measurements associated with aging and general obesity in OSA patients. In addition, adding the face alignment stage provides an increase in uncalibrated-features' performance. Nevertheless, all these handcrafted processes are removed when deep-learning-based features are employed for OSA assessment.
- **Scenario 4:** to our knowledge is the first time that a study on speech- and image- based OSA detection has been carried out on the female population. Consequently, it provides new insights of the relationships between speech, craniofacial characteristics, and OSA on the female population.

As a result of the work described in this Dissertation, several contributions have been published, referring to the study of speech and craniofacial feature in OSA patients.

- **“Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment”.** Espinoza-Cuadros, F., Fernández-Pozo, R., Toledano, D. T., Alcázar-Ramírez, J. D., López-Gonzalo, E. and Hernández-Gómez, L. A. *Computational and Mathematical Methods in Medicine*, vol. 2015, pages 1–13, 2015. ISSN 1748-670X.

This article contains the research of chapters 4 and 6.

- **“Reviewing the connection between speech and obstructive sleep apnea”.** Espinoza-Cuadros, F., Fernández-Pozo, R., Toledano, D. T., Alcázar-Ramírez, J. D., López-Gonzalo, E. and Hernández-Gómez, L. A. *Biomedical engineering online*, vol. 15(1), page 20, 2016. ISSN 1475-925X.

This article contains the research of chapters 4 and 5.

- **“Obstructive Sleep Apnea in Women: Study of Speech and Craniofacial Characteristics”.** Tyan, M., Espinoza-Cuadros, F., Fernández-Pozo, R., Toledano, D. T., López-Gonzalo, E., Alcázar-Ramírez, J. D., and Hernández-Gómez, L. A. *JMIR mHealth and uHealth*, vol. 5(11), page e169, 2017. ISSN 2291-5222.

This article contains the research of chapter 7.

Furthermore, part of the research carried out in this Thesis has also contributed to other works as well as the study of speech disorders for cognitive impairment assessment.

- **“A Spoken Language Database for Research on Moderate Cognitive Impairment: Design and Preliminary Analysis”**. Espinoza-Cuadros, F., Garcia-Zamora, M. A., Torres-Boza, D., Ferrer-Riesgo, C. A., Montero-Benavides, A., Gonzalez-Moreira, E. and Hernández-Gómez, L. A. *In Advances in Speech and Language Technologies for Iberian Languages*, pages 219–228. Springer, 2014.
- **“Automatic Prosodic Analysis to Identify Mild Dementia”**. Gonzalez-Moreira, E., Torres-Boza, D., Kairuz, H. A., Ferrer, C., Garcia-Zamora, M., Espinoza-Cuadros, F. and Hernández-Gómez, L. A. *BioMed Research International*, vol. 2015, pages 1–6, 2015. ISSN 2314-6133.
- **“Formant Frequencies and Bandwidths in Relation to Clinical Variables in an Obstructive Sleep Apnea Population”**. Montero-Benavides, A., Blanco-Murillo, J. L., Fernández-Pozo, R., Espinoza-Cuadros, F., Toledano, D. T., Alcázar-Ramírez, J. D. and Hernández Gómez, L. A. *Journal of Voice*, vol. 30(1), pages 21–29, 2016. ISSN 08921997.

8.4 Future work

A number of future research lines arise from the work conducted in this Thesis. According to each scenario they are detailed below:

- **Scenario 1 and 2:** as regards the speech features (i.e. supervectors, i-vectors), results are limited to the speech from Spanish speakers, so comparison with the other languages will require a more careful analysis of language-dependent acoustic traits in OSA voices. According to ([González-Rodríguez et al., 2014](#)), the i-vector approach is missing relevant information associated with segmental (non-cepstral) and suprasegmental speaker information. Therefore, subspace projection techniques could also be explored over other speech acoustic features previously related to OSA as nasality, voice turbulence or specific co-articulation trajectories. In addition, a comparative analysis of results for both different recordings position, as proposed in ([Kriboy et al., 2014b](#)) should be addressed. In the context of the deep-learning, there is a latent capability of the deep-learning-based techniques to explore new types of “deep” features for speech-based OSA recognition; techniques as transfer learning, as used for OSA face recognition in Chapter 6 and 7. In the same man-

ner, other deep-learning approaches can be explored such as Domain Adversarial Learning and Attentions Models.

- **Scenario 3:** the improvement introduced by the automatic facial alignment stage (i.e. automatic facial landmarking) into the OSA recognition by uncalibrated measurements, encourage to search for alternatives or similar local features that could collect information related to facial phenotyping in OSA subjects without the relationship with clinical features. In addition, good results achieved by deep-learning-based features, encourage to search for similar approaches as other deep network architectures (e.g. deep-learning end-to-end model) which might extract information associated with the facial phenotyping of OSA population removing no OSA-related information.
- **Scenario 4:** besides the search for other sources of speech and craniofacial features to represent the phenotyping in OSA subjects, it would be interesting to go in depth about how hormonal factor could affect the OSA such as the protective mechanisms against apnea episodes and craniofacial differences between different hormonal female groups. Consequently, to evaluate how these changes might affect the speech and craniofacial characterization.

Bibliography

- AHONEN, T., HADID, A. and PIETIKAINEN, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28(12), pages 2037–2041, 2006. ISSN 0162-8828. [106](#)
- AKBANI, R., KWEK, S. and JAPKOWICZ, N. Applying Support Vector Machines to Imbalanced Datasets. In *European conference on machine learning*, pages 39–50. Springer, Berlin, Heidelberg, 2004. ISBN 978-3-540-30115-8. [42](#)
- ALBERT, A. M., RICANEK, K. and PATTERSON, E. A review of the literature on the aging adult skull and face: Implications for forensic science research and applications. *Forensic Science International*, vol. 172(1), pages 1–9, 2007. ISSN 0379-0738. [115](#), [134](#), [150](#), [154](#)
- ALVAR-EZQUERRA, M. El Atlas lingüístico-etnográfico de Andalucía. *Arbor*, vol. 43(157), page 32, 1959. [29](#)
- ÁLVAREZ, D., HORNERO, R., MARCOS, J., DEL-CAMPO, F., ZAMARRÓN, C. and LÓPEZ, M. Applying time, frequency and nonlinear features from nocturnal oximetry to OSA diagnosis. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 3872–3875. IEEE, 2008. [3](#)
- ALVAREZ-CERCADILLO, J., LEE, C. H. and HERNÁNDEZ-GÓMEZ, L. A. Acoustic Modeling of Context Dependent Units, for Large Vocabulary Speech Recognition in Spanish. In *Fourth European Conference on Speech Communication and Technology*. 1995. [35](#)
- AMERICAN ASSOCIATION OF RESPIRATORY CARE-ASSOCIATION OF POLYSOMNOGRAPHY. AARC-APT (American Association of Respiratory Care-Association of Polysomnography Technologists) clinical practice guideline. Polysomnography. *Respiratory care*, vol. 40(12), pages 1336–43, 1995. ISSN 0020-1324. [3](#), [15](#)

- ANTTALAINEN, U., POLO, O., VAHLBERG, T. and SAARES RANTA, T. Women with partial upper airway obstruction are not less sleepy than those with obstructive sleep apnea. *Sleep and Breathing*, vol. 17(2), pages 873–876, 2013. ISSN 1520-9512. [140](#)
- BAHARI, M. H., MCLAREN, M. and VAN-LEE U WEN, D. A. Speaker age estimation using i-vectors. *Engineering Applications of Artificial Intelligence*, vol. 34, pages 99–108, 2014. [37](#), [38](#), [62](#), [68](#), [73](#), [74](#), [75](#), [85](#), [89](#), [93](#), [96](#), [159](#)
- BAHARI, M. H., SAEIDI, R. and VAN-LEE U WEN, D. Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7344–7348. IEEE, 2013. [68](#)
- BAHARI, M. H. and VAN-HAMME, H. Speaker age estimation using Hidden Markov Model weight supervectors. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pages 517–521. 2012. [73](#), [74](#)
- BALAEI, A. T., SUTHERLAND, K., CISTULLI, P. A. and DE CHAZAL, P. Automatic detection of obstructive sleep apnea using facial images. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 215–218. IEEE, 2017. ISBN 978-1-5090-1172-8. [116](#), [117](#), [127](#), [137](#), [138](#)
- BELHUMEUR, P., HESPANHA, J. and KRIEGMAN, D. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(7), pages 711–720, 1997. ISSN 01628828. [106](#)
- BICEGO, M., LAGORIO, A., GROSSO, E. and TISTARELLI, M. On the Use of SIFT Features for Face Authentication. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 35–35. IEEE, 2006. ISBN 0-7695-2646-2. [106](#)
- BIMBOT, F., BONASTRE, J. F., FREDOUILLE, C., GRAVIER, G., MAGRIN-CHAGNOLLEAU, I., MEIGNIER, S., MERLIN, T., ORTEGA-GARCÍA, J., PETROVSKA-DELACRÉTAZ, D. and REYNOLDS, D. A. A tutorial on text-independent speaker verification. *EURASIP journal on applied signal processing*, vol. 2004, pages 430–451, 2004. [62](#), [66](#)
- BIXLER, E. O., VGONTZAS, A. N., LIN, H.-M., TEN HAVE, T., REIN, J., VELA-BUENO, A. and KALES, A. Prevalence of Sleep-disordered Breathing in Women. *American Journal of Respiratory and Critical Care Medicine*, vol. 163(3), pages 608–613, 2001. ISSN 1073-449X. [14](#)

- BIXLER, E. O., VGONTZAS, A. N., TEN-HAVE, T., TYSON, K. and KALES, A. Effects of age on sleep apnea in men: I. Prevalence and severity. *American Journal of Respiratory and Critical Care Medicine*, vol. 157(1), pages 144–148, 1998. [13](#)
- BLANCO-MURILLO, J. L. *Contributions and impact assessment of speech technologies on the automatic detection of severe obstructive sleep apnea syndrome*. PhD thesis, 2013. [24](#), [25](#), [26](#), [27](#)
- BLANCO-MURILLO, J. L., FERNÁNDEZ-POZO, R., LÓPEZ-GONZÁLO, E. and HERNÁNDEZ-GÓMEZ, L. A. Exploring differences between phonetic classes in Sleep Apnoea Syndrome Patients using automatic speech processing techniques. *The Phonetician*, pages 35–54, 2011. [4](#)
- BLAND, J. M. and ALTMAN, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, vol. 327(8476), pages 307–310, 1986. [46](#), [53](#), [54](#)
- BLAND, J. M. and ALTMAN, D. G. Statistics notes. The odds ratio. *BMJ (Clinical research ed.)*, vol. 320(7247), page 1468, 2000. [50](#)
- BLOCK, A., WYNNE, J. W. and BOYSEN, P. G. Sleep-disordered breathing and nocturnal oxygen desaturation in postmenopausal women. *The American Journal of Medicine*, vol. 69(1), pages 75–79, 1980. ISSN 00029343. [141](#)
- BOCKLET, T., MAIER, A., BAUER, J. G., BURKHARDT, F. and NOTH, E. Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1605–1608. IEEE, 2008. [66](#)
- BRADLEY, T. D. Right and left ventricular functional impairment and sleep apnea. *Clinics in chest medicine*, vol. 13(3), pages 459–479, 1992. [11](#)
- BRADLEY, T. D., BROWN, I. G., GROSSMAN, R. F., ZAMEL, N., MARTINEZ, D., PHILLIPSON, E. A. and HOFFSTEIN, V. Pharyngeal size in snorers, nonsnorers, and patients with obstructive sleep apnea. *New England Journal of Medicine*, vol. 315(21), pages 1327–1331, 1986. [12](#)
- BRIDLE, J. S. and BROWN, M. D. An experimental automatic word recognition system. *JSRU Report*, vol. 1003(5), 1974. [65](#)

- BROUGHTON, R., NEVSIMALOVA, S. and ROTH, B. The socio-economic effects of idiopathic hypersomnia—comparison with controls and with compound narcoleptics. *Sleep*, pages 229–233, 1978. [11](#)
- BROWN, I. G., BRADLEY, T. D., PHILLIPSON, E. A., ZAMEL, N. and HOFFSTEIN, V. Pharyngeal Compliance in Snoring Subjects with and without Obstructive Sleep Apnea 1–3. *American Review of Respiratory Disease*, vol. 132(2), pages 211–215, 1985. [21](#)
- BURGES, C. J. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, vol. 2(2), pages 121–167, 1998. ISSN 13845810. [40](#)
- BURGET, L., FAPŠO, M., HUBEIKA, V., GLEMBEK, O., KARAFIÁT, M., KOCKMANN, M., MATĚJKA, P., SCHWARZ, P., HONZA, J. . and SPEECH@FIT, C. BUT system description: NIST SRE 2008. In *NIST Speaker Recognition Evaluation Workshop*. 2008. [90](#)
- BURGOS-ARTIZZU, X. P., PERONA, P. and DOLLAR, P. Robust Face Landmark Estimation under Occlusion. In *2013 IEEE International Conference on Computer Vision*, pages 1513–1520. IEEE, Sydney, NSW, Australia, 2013. ISBN 978-1-4799-2840-8. [107](#), [125](#)
- BUXTON, O. M. and MARCELLI, E. Short and long sleep are positively associated with obesity, diabetes, hypertension, and cardiovascular disease among adults in the United States. *Social science & medicine*, vol. 71(5), pages 1027–1036, 2010. [1](#)
- CAKIRER, B., HANS, M. G., GRAHAM, G., AYLOR, J., TISHLER, P. V. and REDLINE, S. The relationship between craniofacial morphology and obstructive sleep apnea in whites and in African-Americans. *American journal of respiratory and critical care medicine*, vol. 163(4), pages 947–50, 2001. ISSN 1073-449X. [9](#), [103](#)
- CAMPBELL, W. M., CAMPBELL, J. P., REYNOLDS, D. A., SINGER, E. and TORRES-CARRASQUILLO, P. A. Support vector machines for speaker and language recognition. *Computer Speech and Language*, vol. 20(2-3 SPEC. ISS.), pages 210–229, 2006a. ISSN 08852308. [62](#), [69](#)
- CAMPBELL, W. M., STURIM, D. E. and REYNOLDS, D. A. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, vol. 13(5), pages 308–311, 2006b. ISSN 10709908. [39](#), [69](#)
- CERRATO, L., FALCONE, M. and PAOLONI, A. Subjective age estimation of telephonic voices. *Speech Communication*, vol. 31(2), pages 107–112, 2000. [74](#)

- CHANG, C.-C. and LIN, C.-J. LIBSVM. *ACM Transactions on Intelligent Systems and Technology*, vol. 2(3), pages 1–27, 2011. ISSN 21576904. [75](#)
- CHELLAPPA, R., WILSON, C. and SIROHEY, S. Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, vol. 83(5), pages 705–741, 1995. ISSN 00189219. [105](#)
- CHEN, D., CAO, X., WANG, L., WEN, F. and SUN, J. Bayesian Face Revisited: A Joint Formulation. pages 566–579. Springer, Berlin, Heidelberg, 2012. [128](#)
- CHERVIN, R. D. Sleepiness, Fatigue, Tiredness, and Lack of Energy in Obstructive Sleep Apnea. *Chest*, vol. 118(2), pages 372–379, 2000. ISSN 00123692. [13](#)
- CHEUNG, K., ISHMAN, S. L., BENKE, J. R., COLLOP, N., TRON, L., MOY, N. and STIERER, T. L. Prediction of obstructive sleep apnea using visual photographic analysis. *Journal of Clinical Anesthesia*, vol. 32, pages 40–46, 2016. ISSN 09528180. [104](#)
- CHI, L., COMYN, F.-L., MITRA, N., REILLY, M. P., WAN, F., MAISLIN, G., CHMIEWSKI, L., THORNE-FITZGERALD, M. D., VICTOR, U. N., PACK, A. I. and SCHWAB, R. J. Identification of craniofacial risk factors for obstructive sleep apnoea using three-dimensional MRI. *European Respiratory Journal*, vol. 38(2), pages 348–358, 2011. ISSN 0903-1936. [117](#)
- CHUNG, F., YEGNESWARAN, B., LIAO, P., CHUNG, S. A., VAIRAVANATHAN, S., ISLAM, S., KHAJEHDEHI, A. and SHAPIRO, C. M. STOP Questionnaire. *Anesthesiology*, vol. 108(5), pages 812–821, 2008. ISSN 0003-3022. [2](#), [15](#)
- COOTES, T., EDWARDS, G. and TAYLOR, C. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(6), pages 681–685, 2001. ISSN 01628828. [103](#), [106](#), [107](#), [108](#), [118](#), [120](#), [125](#)
- COOTES, T., TAYLOR, C., COOPER, D. and GRAHAM, J. Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, vol. 61(1), pages 38–59, 1995. ISSN 10773142. [107](#), [118](#), [119](#), [122](#)
- CORTES, C. and VAPNIK, V. Support-vector networks. *Machine Learning*, vol. 20(3), pages 273–297, 1995. ISSN 0885-6125. [39](#)
- CYRAN, K. A., KAWULOK, J., KAWULOK, M., STAWARZ, M., MICHALAK, M., PIETROWSKA, M., WIDŁAK, P. and POLAŃSKA, J. Support Vector Machines in Biomedical and Biometrical Applications. pages 379–417. Springer, Berlin, Heidelberg, 2013. [44](#)

- DANCEY, D. R., HANLY, P. J., SOONG, C., LEE, B. and HOFFSTEIN, V. Impact of menopause on the prevalence and severity of sleep apnea. *Chest*, vol. 120(1), pages 151–5, 2001. ISSN 0012-3692. [140](#)
- DARLEY, F. L., ARONSON, A. E. and BROWN, J. R. Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, vol. 12(2), pages 246–69, 1969. ISSN 0022-4685. [19](#)
- DAVIDSON, T. M. The Great Leap Forward: the anatomic basis for the acquisition of speech and obstructive sleep apnea. *Sleep Medicine*, vol. 4(3), pages 185–194, 2003. ISSN 13899457. [3](#), [12](#), [16](#)
- DEHAK, N. *Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification*. PhD thesis, Ecole de Technologie Superieure (Canada), 2009. [71](#)
- DEHAK, N., KENNY, P. J., DEHAK, R., DUMOUCHEL, P. and OUELLET, P. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19(4), pages 788–798, 2011a. ISSN 1558-7916. [62](#), [65](#), [68](#), [71](#), [73](#), [76](#), [159](#)
- DEHAK, N., TORRES-CARRASQUILLO, P. A., REYNOLDS, D. A. and DEHAK, R. Language Recognition via i-vectors and Dimensionality Reduction. In *Interspeech*, pages 857–860. 2011b. [68](#)
- DEMENT, W. C. and LEARY, E. B. Sleep disorders: A widely ignored pandemic. *Journal for Respiratory Care & Sleep Medicine*, page 2, 2009. [14](#)
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. [57](#)
- DINGES, D. F. An overview of sleepiness and accidents. *Journal of sleep research*, vol. 4(S2), pages 4–14, 1995. ISSN 1365-2869. [1](#)
- DOBRY, G., HECHT, R. M., AVIGAL, M. and ZIGEL, Y. Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19(7), pages 1975–1985, 2011. ISSN 1558-7916. [73](#), [74](#)

- DRAPER, N. R. and SMITH, H. *Applied Regression Analysis*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1998. ISBN 9781118625590. [143](#)
- DURÁN-CANTOLLA, J. Necessity to treat also affected people with mild or moderate apnoea. In *3rd International Congress on Sleep Apnoea*. 2012. [24](#), [32](#)
- DURÁN-CANTOLLA, J., MAR, J., DE LA TORRE MUÑECAS, G., RUBIO ARAMENDI, R. and GUERRA, L. The Availability in Spanish Public Hospitals of Resources for Diagnosing and Treating Sleep Apnea-Hypopnea Syndrome. *Archivos de Bronconeumología ((English Edition))*, vol. 40(6), pages 259–267, 2004. ISSN 15792129. [3](#)
- EL-RAI, M. C., WERGI, N., AL-MUHAIRI, H. and ALSAFAR, H. Using facial images for the diagnosis of genetic syndromes: A survey. In *2015 International Conference on Communications, Signal Processing, and their Applications (ICCSPA '15)*, pages 1–6. IEEE, IEEE, 2015. [2](#), [105](#)
- ESPINOZA-CUADROS, F., FERNÁNDEZ-POZO, R., TOLEDANO, D. T., ALCÁZAR-RAMÍREZ, J. D., LÓPEZ-GONZALO, E. and HERNÁNDEZ-GÓMEZ, L. A. Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment. *Computational and Mathematical Methods in Medicine*, vol. 2015, pages 1–13, 2015. ISSN 1748-670X. [31](#), [83](#), [108](#)
- ESPINOZA-CUADROS, F., FERNÁNDEZ-POZO, R., TOLEDANO, D. T., ALCÁZAR-RAMÍREZ, J. D., LÓPEZ-GONZALO, E. and HERNÁNDEZ-GÓMEZ, L. A. Reviewing the connection between speech and obstructive sleep apnea. *Biomedical engineering online*, vol. 15(1), page 20, 2016. ISSN 1475-925X. [24](#), [31](#), [83](#), [96](#)
- ESPINOZA-CUADROS, F., GARCIA-ZAMORA, M. A., TORRES-BOZA, D., FERRER-RIESGO, C. A., MONTERO-BENAVIDES, A., GONZALEZ-MOREIRA, E. and HERNANDEZ-GÓMEZ, L. A. A Spoken Language Database for Research on Moderate Cognitive Impairment: Design and Preliminary Analysis. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 219–228. Springer, 2014. [2](#)
- ESTEVA, A., KUPREL, B., NOVOA, R. A., KO, J., SWETTER, S. M., BLAU, H. M. and THRUN, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, vol. 542(7639), pages 115–118, 2017. ISSN 0028-0836. [128](#)
- EVANS, S., NEAVE, N. and WAKELIN, D. Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology*, vol. 72(2), pages 160–163, 2006. ISSN 03010511. [75](#)

- FAIRBANKS, G. The rainbow passage. *Voice and articulation drillbook*, vol. 2, 1960. [19](#)
- FAWCETT, T. ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Laboratories, Palo Alto, CA 94304, 2004. [52](#)
- FERNÁNDEZ-POZO, R. *Técnicas de caracterización de voces como apoyo a la detección de la apnea del sueño*. PhD thesis, 2011. [5](#), [24](#), [27](#)
- FERNÁNDEZ-POZO, R., BLANCO-MURILLO, J. L., HERNÁNDEZ-GÓMEZ, L. A., LÓPEZ-GONZÁLO, E., ALCÁZAR-RAMÍREZ, J. and TOLEDANO, D. T. Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker Recognition Techniques. *EURASIP Journal on Advances in Signal Processing*, vol. 2009(1), page 982531, 2009. ISSN 1687-6180. [2](#), [3](#), [4](#), [21](#), [24](#), [62](#), [63](#), [66](#), [68](#), [91](#), [96](#), [97](#), [98](#)
- FERNÁNDEZ-POZO, R., HERNÁNDEZ-GÓMEZ, L. A., LÓPEZ-GONZÁLO, E., ALCÁZAR-RAMÍREZ, J., PORTILLO, G. and TOLEDANO, D. T. Design of a multimodal database for research on automatic detection of severe apnoea cases. 2008. [23](#), [25](#)
- FISCHER, Y., JUNGE-HÜLSING, B., RETTINGER, G. and PANIS, A. The use of an ambulatory, automatic sleep recording device (QUISITM Version 1.0) in the evaluation of primary snoring and obstructive sleep apnoea. *Clinical Otolaryngology and Allied Sciences*, vol. 29(1), pages 18–23, 2004. [16](#)
- FIZ, J. A., MORERA, J., ABAD, J., BELSUNCES, A., HARO, M., FIZ, J. I., JANE, R., CAMINAL, P. and RODENSTEIN, D. Acoustic analysis of vowel emission in obstructive sleep apnea. *Chest*, vol. 104(4), pages 1093–6, 1993. ISSN 0012-3692. [20](#), [26](#), [61](#)
- FLEMONS, W. W., BUYSSE, D., REDLINE, S., OACK, A., STROHL, K., WHEATLEY, J., YOUNG, T., DOUGLAS, N., LEVY, P., MCNICOLAS, W., FLEETHAM, J., WHITE, D., SCHMIDT-NOWARRA, W., CARLEY, D. and ROMANIUK, J. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force. *Sleep*, vol. 22(5), pages 667–89, 1999. ISSN 0161-8105. [15](#)
- FLEMONS, W. W., LITTNER, M. R., ROWLEY, J. A., GAY, P., ANDERSON, W. M., HUDGEL, D. W., MCEVOY, R. D. and LOUBE, D. I. Home Diagnosis of Sleep Apnea: A Systematic Review of the Literature. *Chest*, vol. 124(4), pages 1543–1579, 2003. ISSN 00123692. [3](#)

- FOSTER, K. R., KOPROWSKI, R. and SKUFCA, J. D. Machine learning, medical diagnosis, and biomedical engineering research - commentary. *BioMedical Engineering OnLine*, vol. 13(1), page 94, 2014. ISSN 1475-925X. [5](#), [44](#), [45](#), [86](#), [93](#), [97](#), [100](#), [102](#)
- FOWLER, J., COHEN, L. and JARVIS, P. *Practical statistics for field biology*. John Wiley & Sons, 1998. ISBN 9781118685648. [55](#), [78](#), [94](#)
- FOX, A. W., MONOSON, P. K. and MORGAN, C. D. Speech Dysfunction of Obstructive Sleep Apnea. *Chest*, vol. 96(3), pages 589–595, 1989. ISSN 00123692. [5](#), [19](#), [20](#), [23](#), [25](#), [26](#), [27](#), [28](#), [61](#)
- FURUI, S. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29(2), pages 254–272, 1981. ISSN 0096-3518. [34](#)
- GAUVAIN, J. L. and LEE, C. H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE transactions on speech and audio processing*, vol. 2(2), pages 291–298, 1994. [57](#)
- GLAS, A. S., LIJMER, J. G., PRINS, M. H., BONSEL, G. J. and BOSSUYT, P. M. M. The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*, vol. 56(11), pages 1129–1135, 2003. [50](#)
- GODINO-LLORENTE, J. I., GOMEZ-VILDA, P. and BLANCO-VELASCO, M. Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. *IEEE transactions on biomedical engineering*, vol. 53(10), pages 1943–1953, 2006. [66](#)
- GOLDSHTEIN, E., TARASIUK, A. and ZIGEL, Y. Automatic detection of obstructive sleep apnea using speech signals. *IEEE transactions on bio-medical engineering*, vol. 58(5), pages 1373–82, 2011. ISSN 1558-2531. [21](#), [62](#), [63](#), [66](#), [68](#), [91](#), [96](#), [98](#), [144](#), [147](#), [158](#)
- GÓMEZ-VILDA, P., RODELLAR-BIARGE, V., NIETO-LLUIS, V., MUÑOZ-MULAS, C., MAZAIRA-FERNÁNDEZ, L. M., RAMÍREZ-CALVO, C., FERNÁNDEZ-FERNÁNDEZ, M. and TORIBIO-DÍAZ, E. Neurological disease detection and monitoring from voice production. In *NOLISP'11 Proceedings of the 5th international conference on Advances in nonlinear speech processing* (edited by Springer-Verlag Berlin), pages 1–8. Las Palmas de Gran Canaria, Spain, 2011. [2](#)
- GONZÁLEZ-DOMÍNGUEZ, J. *Session Variability Compensation in Automatic Speaker and Language Recognition*. PhD thesis, Universidad Autónoma de Madrid, 2011. [71](#)

- GONZÁLEZ-RODRÍGUEZ, J., GIL, J., PÉREZ, R. and FRANCO-PEDROSO, J. What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials. In *Proceedings of Odyssey*, pages 33–40, 2014. [85](#), [91](#), [163](#)
- GUIDRY, U. C., MENDES, L. A., EVANS, J. C., LEVY, D., O’CONNOR, G. T., LARSON, M. G., GOTTLIEB, D. J. and BENJAMIN, E. J. Echocardiographic Features of the Right Heart in Sleep-Disordered Breathing. *American Journal of Respiratory and Critical Care Medicine*, vol. 164(6), pages 933–938, 2001. ISSN 1073-449X. [11](#)
- GUILLEMINAULT, C., QUERA-SALVA, M.-A., PARTINEN, M. and JAMIESON, A. Women and the Obstructive Sleep Apnea Syndrome. *Chest*, vol. 93(1), pages 104–109, 1988. ISSN 00123692. [100](#), [117](#), [149](#), [151](#)
- GUILLEMINAULT, C., RILEY, R. and POWELL, N. Obstructive Sleep Apnea and Abnormal Cephalometric Measurements. *Chest*, vol. 86(5), pages 793–794, 1984. ISSN 00123692. [2](#), [3](#), [12](#), [17](#), [20](#), [61](#), [103](#), [112](#), [149](#)
- GUILLEMINAULT, C., TILKIAN, A. and DEMENT, W. C. The sleep apnea syndromes. *Annual review of medicine*, vol. 27(1), pages 465–84, 1976. ISSN 0066-4219. [9](#)
- HAMDAN, A., AL-BARAZI, R., KHNEIZER, G., TURFE, Z., SINNO, S., ASHKAR, J. and TABRI, D. Formant frequency in relation to body mass composition. *Journal of voice : official journal of the Voice Foundation*, vol. 27(5), pages 567–71, 2013. [75](#)
- HEDNER, J. A., WILCOX, I. and SULLIVAN, C. E. Speculations on the interaction between vascular disease and obstructive sleep apnea. *Lung biology in health and disease*, vol. 71, pages 823–846, 1994. [11](#)
- HEINZER, R. C., STANCHINA, M. L., MALHOTRA, A., JORDAN, A. S., PATEL, S. R., LO, Y.-L., WELLMAN, A., SCHORY, K., DOVER, L. and WHITE, D. P. Effect of increased lung volume on sleep disordered breathing in patients with sleep apnoea. *Thorax*, vol. 61(5), pages 435–439, 2006. ISSN 0040-6376. [10](#)
- HOFFSTEIN, V. and SZALAI, J. P. Predictive Value of Clinical Features in Diagnosing Obstructive Sleep Apnea. *Sleep*, vol. 16(2), pages 118–22, 1993. ISSN 1550-9109. [15](#), [82](#), [98](#), [101](#), [143](#), [144](#)
- HORNE, J. A. and REYNER, L. A. Driver sleepiness. *Journal of Sleep Research*, vol. 4(S2), pages 23–29, 1995. ISSN 09621105. [1](#), [11](#)

- HSU, C. W., CHANG, C. C. and LIN, C. J. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2003. [45](#)
- HUANG, X., ACERO, A. and HON, H.-W. W. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001. ISBN 0130226165. [101](#)
- JAMIESON, A., GUILLEMINAULT, C., PARTINEN, M. and QUERA-SALVA, M. A. Obstructive Sleep Apneic Patients Have Craniomandibular Abnormalities. *Sleep*, vol. 9(4), pages 469–477, 1986. ISSN 0161-8105. [17](#), [112](#)
- JIN, X. and TAN, X. Face alignment in-the-wild: A Survey. *Computer Vision and Image Understanding*, vol. 162, pages 1–22, 2017. ISSN 10773142. [107](#)
- JOHNS, M. W. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *sleep*, vol. 14(6), pages 540–545, 1991. [15](#)
- JORDAN, A. S., WELLMAN, A., EDWARDS, J. K., SCHORY, K., DOVER, L., MACDONALD, M., PATEL, S. R., FOGEL, R. B., MALHOTRA, A. and WHITE, D. P. Respiratory control stability and upper airway collapsibility in men and women with obstructive sleep apnea. *Journal of Applied Physiology*, vol. 99(5), pages 2020–2027, 2005. ISSN 8750-7587. [140](#)
- KASS, M., WITKIN, A. and TERZOPOULOS, D. Snakes: Active contour models. *International Journal of Computer Vision*, vol. 1(4), pages 321–331, 1988. ISSN 0920-5691. [107](#)
- KENNY, P., OUELLET, P., DEHAK, N., GUPTA, V. and DUMOUCHEL, P. A Study of Interspeaker Variability in Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16(5), pages 980–988, 2008. ISSN 1558-7916. [55](#), [73](#)
- KINNUNEN, T. and LI, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, vol. 52(1), pages 12–40, 2010. ISSN 01676393. [66](#)
- KITANO, I., PARK, S., KATO, K., NITTA, N., TAKATO, T. and SUSAMI, T. Craniofacial Morphology of Conotruncal Anomaly Face Syndrome. *The Cleft Palate-Craniofacial Journal*, vol. 34(5), pages 425–429, 1997. ISSN 1055-6656. [2](#)
- KOLLIAS, I. Adult craniocervical and pharyngeal changes - a longitudinal cephalometric study between 22 and 42 years of age. Part 1: morphological craniocervical and hyoid bone changes. *The European Journal of Orthodontics*, vol. 21(4), pages 333–344, 1999. ISSN 0141-5387. [149](#)

- KOLLIAS, I. and KROGSTAD, O. Adult craniofacial and pharyngeal changes - a longitudinal cephalometric study between 22 and 42 years of age. Part II: morphological uvuloglossopharyngeal changes. *The European Journal of Orthodontics*, vol. 21(4), pages 345–355, 1999. ISSN 0141-5387. [149](#)
- KRIBOY, M., TARASIUK, A. and ZIGEL, Y. Obstructive sleep apnea detection using speech signals. In *Proceedings of the annual conference of the Afeka-AVIOs in Speech Processing*, pages 1–5. 2013. [96](#), [98](#)
- KRIBOY, M., TARASIUK, A. and ZIGEL, Y. A novel method for obstructive sleep apnea severity estimation using speech signals. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3606–3610. IEEE, IEEE, 2014a. ISBN 978-1-4799-2893-4. [80](#), [81](#), [95](#), [96](#), [100](#), [101](#)
- KRIBOY, M., TARASIUK, A. and ZIGEL, Y. Detection of Obstructive sleep apnea in awake subjects by exploiting body posture effects on the speech signal. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4224–4227. IEEE, IEEE, 2014b. ISBN 978-1-4244-7929-0. [62](#), [91](#), [99](#), [163](#)
- KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, vol. 60, pages 1097–1105. ACM, 2012. ISSN 00010782. [129](#)
- KUSHIDA, C. A. *Handbook of sleep disorders*. CRC Press, 2008. ISBN 9780849373190. [1](#)
- KUSHIDA, C. A., EFRON, B. and GUILLEMINAULT, C. A predictive morphometric model for the obstructive sleep apnea syndrome. *Annals of Internal Medicine*, vol. 127(8.Part.1), pages 581–587, 1997. [15](#)
- KWON, Y. H. and LOBO, N. D. V. Age classification from facial images. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 762–767. IEEE, 1994. [5](#)
- LAM, B., IP, M. S. M., TENCH, E. and RYAN, C. F. Craniofacial profile in Asian and white subjects with obstructive sleep apnoea. *Thorax*, vol. 60(6), pages 504–510, 2005. [13](#)
- LAM, J. C. M., SHARMA, S. K. and LAM, B. Obstructive sleep apnoea: definitions, epidemiology & natural history. *The Indian journal of medical research*, vol. 131, pages 165–70, 2010. ISSN 0971-5916. [1](#), [2](#), [10](#), [14](#), [87](#)

- LEE, B. J., KU, B., JANG, J. and KIM, J. A novel method for classifying body mass index on the basis of speech signals for future clinical applications: a pilot study. *Evidence-based complementary and alternative medicine : eCAM*, vol. 2013, page 150265, 2013. ISSN 1741-427X. [62](#), [66](#), [75](#)
- LEE, R. W. W., CHAN, A. S. L., GRUNSTEIN, R. R. and CISTULLI, P. A. Craniofacial phenotyping in obstructive sleep apnea—a novel quantitative photographic approach. *Sleep*, vol. 32(1), pages 37–45, 2009a. ISSN 0161-8105. [4](#), [5](#), [7](#), [17](#), [103](#), [104](#), [106](#), [108](#), [110](#), [111](#), [112](#), [114](#), [115](#), [116](#), [117](#), [121](#), [128](#), [136](#), [137](#), [138](#), [150](#), [158](#), [162](#)
- LEE, R. W. W., PETOCZ, P., PRVAN, T., CHAN, A. S. L., GRUNSTEIN, R. R. and CISTULLI, P. A. Prediction of Obstructive Sleep Apnea with Craniofacial Photographic Analysis. *Sleep*, vol. 32(1), pages 46–52, 2009b. ISSN 1550-9109. [103](#), [110](#), [116](#), [117](#), [118](#), [127](#), [137](#), [158](#), [160](#)
- LEE, R. W. W., SUTHERLAND, K., CHAN, A. S. L., ZENG, B., GRUNSTEIN, R. R., DARENDELILER, M. A., SCHWAB, R. J. and CISTULLI, P. A. Relationship Between Surface Facial Dimensions and Upper Airway Structures in Obstructive Sleep Apnea. *Sleep*, vol. 33(9), pages 1249–1254, 2010. ISSN 1550-9109. [5](#), [18](#), [103](#), [104](#), [106](#), [108](#), [111](#), [134](#), [153](#), [154](#), [158](#)
- LEECH, J. A., ÖNAL, E., DULBERG, C. and LOPATA, M. A. A Comparison of Men and Women with Occlusive Sleep Apnea Syndrome. *Chest*, vol. 94(5), pages 983–988, 1988. ISSN 00123692. [150](#)
- LIN, C. M., DAVIDSON, T. M. and ANCOLI-ISRAEL, S. Gender differences in obstructive sleep apnea and treatment implications. *Sleep Medicine Reviews*, vol. 12(6), pages 481–496, 2008. ISSN 10870792. [140](#), [149](#)
- LINDE, Y., BUZO, A. and GRAY, R. An algorithm for vector quantizer design. *IEEE Transactions on communications*, vol. 28(1), pages 84–95, 1980. [57](#)
- LLOBERES, P., LEVY, G., DESCALS, C., SAMPOL, G., ROCA, A., SAGALES, T. and DE-LA CALZADA, M. D. Self-reported sleepiness while driving as a risk factor for traffic accidents in patients with obstructive sleep apnoea syndrome and in non-apnoeic snorers. *Respiratory medicine*, vol. 94(10), pages 971–976, 2000. [2](#)
- LÓPEZ-GONZÁLO, E., CAMINERO, J., CORTÁZAR, I. and HERNÁNDEZ-GÓMEZ, L. A. Improvement on connected numbers recognition using prosodic information. In *ICSLP*. 1998. [35](#)

- LOUBE, I. M. D., POCETA, J. S., MORALES, M. C., PEACOCK, D. M. M. and MITLER, M. M. Self-reported Snoring in Pregnancy. *Chest*, vol. 109(4), pages 885–889, 1996. ISSN 00123692. [14](#)
- LOWE, A. A., FLEETHAM, J. A., ADACHI, S. and RYAN, C. Cephalometric and computed tomographic predictors of obstructive sleep apnea severity. *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 107(6), pages 589–595, 1995. ISSN 08895406. [2](#), [3](#), [12](#), [17](#), [103](#), [112](#)
- LYBERG, T., KROGSTAD, O. and DJUPESLAND, G. Cephalometric analysis in patients with obstructive sleep apnoea syndrome. I. Skeletal morphology. *The Journal of laryngology and otology*, vol. 103(3), pages 287–92, 1989. ISSN 0022-2151. [17](#), [103](#)
- MAISLIN, G., PACK, A. I., KRIBBS, N. B., SMITH, P. L., SCHWARTZ, A. R., KLINE, L. R., SCHWAB, R. J. and DINGES, D. F. A Survey Screen for Prediction of Apnea. *Sleep*, vol. 18(3), pages 158–166, 1995. ISSN 1550-9109. [15](#)
- MALHOTRA, A., HUANG, Y., FOGEL, R., LAZIC, S., PILLAR, G., JAKAB, M., KIKINIS, R. and WHITE, D. P. Aging Influences on Pharyngeal Anatomy and Physiology: The Predisposition to Pharyngeal Collapse. *The American Journal of Medicine*, vol. 119(1), pages 9–72, 2006. ISSN 00029343. [13](#)
- MALLAMPATI, S. R., GATT, S. P., GUGINO, L. D., DESAI, S. P., WARAKSA, B., FREIBERGER, D. and LIU, P. L. A clinical sign to predict difficult tracheal intubation; a prospective study. *Canadian Anaesthetists' Society Journal*, vol. 32(4), pages 429–434, 1985. ISSN 0008-2856. [2](#), [15](#)
- MARTINEZ, F., TAPIAS, D., ALVAREZ, J. and LEON, P. Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In *Fifth European Conference on Speech Communication and Technology*. 1997. [36](#)
- MONOSON, P. K. and FOX, A. W. Preliminary observation of speech disorder in obstructive and mixed sleep apnea. *Chest*, vol. 92(4), pages 670–5, 1987. ISSN 0012-3692. [3](#), [18](#), [19](#), [20](#), [23](#), [25](#), [27](#), [28](#), [61](#)
- MONTERO-BENAVIDES, A. *Statistical analysis of the connection between sleep apnea and speech*. PhD thesis, Universidad Politécnica de Madrid, 2017. [24](#)
- MONTERO-BENAVIDES, A., BLANCO-MURILLO, J. L., FERNÁNDEZ, A., FERNÁNDEZ-POZO, R., TOLEDANO, D. T. and HERNÁNDEZ-GÓMEZ, L. A. Using HMM to detect speakers with

- severe obstructive sleep apnoea syndrome. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 121–128. Springer, 2012. [62](#), [66](#), [96](#), [97](#), [98](#)
- MONTERO-BENAVIDES, A., BLANCO-MURILLO, J. L., FERNÁNDEZ-POZO, R., ESPINOZA-CUADROS, F., TOLEDANO, D. T., ALCÁZAR-RAMÍREZ, J. D. and HERNÁNDEZ GÓMEZ, L. A. Formant Frequencies and Bandwidths in Relation to Clinical Variables in an Obstructive Sleep Apnea Population. *Journal of Voice*, vol. 30(1), pages 21–29, 2016. ISSN 08921997. [4](#), [21](#), [24](#), [31](#), [63](#), [96](#), [97](#)
- MONTERO-BENAVIDES, A., FERNÁNDEZ-POZO, R., TOLEDANO, D. T., BLANCO-MURILLO, J. L., LÓPEZ-GONZÁLO, E. and HERNÁNDEZ-GÓMEZ, L. A. Analysis of voice features related to obstructive sleep apnoea and their application in diagnosis support. *Computer Speech & Language*, vol. 28(2), pages 434–452, 2014. ISSN 08852308. [62](#), [63](#), [85](#), [99](#)
- MORENO-BILBAO, M. A., POIG, D., BONAFONTE-CÁVEZ, A., LLEIDA, E., LLISTERRI, J., MARIÑO-ACEBAL, J. B. and NADEU-CAMPRUBÍ, C. Albayzin speech database: Design of the phonetic corpus. In *EUROSPEECH, 1993.*, pages 175–178. . EUROSPEECH, 1993. [28](#), [35](#)
- MORTIMORE, I. L., MARSHALL, I., WRAITH, P. K., SELLAR, R. J. and DOUGLAS, N. J. Neck and total body fat deposition in nonobese and obese patients with sleep apnea compared with that in control subjects. *American journal of respiratory and critical care medicine*, vol. 157(1), pages 280–3, 1998. ISSN 1073-449X. [110](#)
- MURILLO-BLANCO, J. L., HERNÁNDEZ-GÓMEZ, L. A., FERNÁNDEZ-POZO, R. and RAMOS-CASTRO, D. Improving automatic detection of obstructive sleep apnea through nonlinear analysis of sustained speech. *Cognitive Computation*, vol. 5(4), pages 458–472, 2013. [91](#)
- NAIK, J., NETSCH, L. and DODDINGTON, G. Speaker verification over long distance telephone lines. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 524–527. IEEE, 1989. [35](#)
- NATALE, V., DREJAK, M., ERBACCI, A., TONETTI, L., FABBRI, M. and MARTONI, M. Monitoring sleep with a smartphone accelerometer. *Sleep and Biological Rhythms*, vol. 10(4), pages 287–292, 2012. [16](#)
- NETZER, N. C., STOOHS, R. A., NETZER, C. M., CLARK, K. and STROHL, K. P. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. *Annals of internal medicine*, vol. 131(7), pages 485–491, 1999. [15](#)

- ONG, K. C. and CLERK, A. A. Comparison of the severity of sleep-disordered breathing in Asian and Caucasian patients seen at a sleep disorders center. *Respiratory medicine*, vol. 92(6), pages 843–848, 1998. [14](#)
- OSUNA, E., FREUND, R. and GIROSIT, F. Training support vector machines: an application to face detection. In *Computer vision and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on*, pages 130–136. IEEE, 1997. [39](#)
- PAN, S. J. and YANG, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22(10), pages 1345–1359, 2010. ISSN 1041-4347. [129](#)
- PARKHI, O. M., VEDALDI, A. and ZISSERMAN, A. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference 2015*, pages 1–41. British Machine Vision Association, 2015. ISBN 1-901725-53-7. [106](#), [128](#), [131](#), [152](#)
- PAYNE, R. J., KOST, K. M., FRENKIEL, S., ZEITOUNI, A. G., SEJEAN, G., SWEET, R. C., NAOR, N., HERNANDEZ, L. and KIMOFF, R. J. Laryngeal inflammation assessed using the reflux finding score in obstructive sleep apnea. *Otolaryngology-Head and Neck Surgery*, vol. 134(5), pages 836–842, 2006. [27](#)
- PELECANOS, J., CHAUDHARI, U. and RAMASWAMY, G. Compensation of utterance length for speaker verification. In *ODYSSEY04-The Speaker and Language Recognition Workshop*. 2004. [67](#)
- PELLOM, B. L. and HANSEN, J. H. L. Voice analysis in adverse conditions: the centennial Olympic park bombing 911 call. In *Circuits and Systems, 1997. Proceedings of the 40th Midwest Symposium on*, vol. 2, pages 873–876. IEEE, 1997. [66](#)
- PEPPARD, P. E., YOUNG, T., BARNET, J. H., PALTA, M., HAGEN, E. W. and HLA, K. M. Increased Prevalence of Sleep-Disordered Breathing in Adults. *American Journal of Epidemiology*, vol. 177(9), pages 1006–1014, 2013. ISSN 0002-9262. [2](#), [9](#), [139](#)
- PETERSON-FALZONE, S. J. Resonance disorders in structural defects. *N. Lass, L*, 1982. [19](#)
- POORJAM, A. H., BAHARI, M. H. and VAN HAMME, H. Speaker weight estimation from speech signals using a fusion of the i-vector and NFA frameworks. In *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pages 118–123. IEEE, 2015a. ISBN 978-1-4799-8818-1. [5](#), [37](#), [38](#), [73](#), [74](#), [75](#), [93](#), [96](#)

- POORJAM, A. H., BAHARI, M. H., VASILAKAKIS, V., VAN HAMME, H. and VAN-HAMME, H. Height estimation from speech signals using i-vectors and least-squares support vector regression. In *Telecommunications and Signal Processing (TSP), 2015 38th International Conference on*, pages 1–5. IEEE, IEEE, 2015b. ISBN 978-1-4799-8498-5. [37](#), [38](#), [62](#), [68](#), [73](#), [74](#), [75](#), [93](#), [96](#)
- PREMKUMAR, R., ARUN, C. and DIVYA, R. S. Continuous Wavelet Analysis of the Breathing Sounds of Obstructive Sleep Apnea Patients and Normal Subjects during Wakefulness. *Applied Mechanics and Materials*, vol. 622(622), pages 45–50, 2014. ISSN 1662-7482. [91](#)
- PUERTAS, F. J., PIN, G., MARÍA, J. M. and DURÁN, J. Documento de consenso Nacional sobre el síndrome de Apneas-hipopneas del sueño. *Grupo Español De Sueño*, vol. 164, 2005. [1](#), [10](#), [12](#), [14](#), [16](#)
- PUNJABI, N. M. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, vol. 5(2), pages 136–43, 2008. ISSN 1546-3222. [10](#), [11](#), [14](#)
- RAZAVIAN, A. S., AZIZPOUR, H., SULLIVAN, J. and CARLSSON, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519. IEEE, 2014. ISBN 978-1-4799-4308-1. ISSN 2160-7516. [106](#), [129](#), [130](#), [131](#)
- REYES, B. A., RELJIN, N. and CHON, K. H. Tracheal sounds acquisition using smartphones. *Sensors*, vol. 14(8), pages 13830–13850, 2014. [16](#)
- REYNOLDS, D. A., QUATIERI, T. F. and DUNN, R. B. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, vol. 10(1-3), pages 19–41, 2000. ISSN 10512004. [35](#), [55](#), [57](#), [58](#), [67](#)
- ROBB, M. P., YATES, J. and MORGAN, E. J. Vocal Tract Resonance Characteristics of Adults with Obstructive Sleep Apnea. *Acta Oto-Laryngologica*, vol. 117(5), pages 760–763, 1997. ISSN 0001-6489. [5](#), [20](#), [21](#), [26](#), [27](#), [28](#), [31](#), [61](#), [63](#), [96](#)
- RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATY, A., KHOSLA, A., BERNSTEIN, M., BERG, A. C. and FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, vol. 115(3), pages 211–252, 2015. ISSN 0920-5691. [127](#)
- RYAN, C. M. and BRADLEY, T. D. Pathogenesis of obstructive sleep apnea. *Journal of Applied Physiology*, vol. 99(6), pages 2440–2450, 2005. [10](#), [12](#)

- SADJADI, S. O., SLANEY, M. and HECK, L. MSR identity toolbox v1. 0: A matlab toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter*, vol. 1(4), 2013. [69](#)
- SCHELLENBERG, J. B., MAISLIN, G. and SCHWAB, R. J. Physical findings and the risk for obstructive sleep apnea. The importance of oropharyngeal structures. *American journal of respiratory and critical care medicine*, vol. 162(2 Pt 1), pages 740–8, 2000. ISSN 1073-449X. [9](#)
- SCHÖLKOPF, B. and SMOLA, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002. [43](#)
- SCHWAB, R. J., PASIRSTEIN, M., PIERSON, R., MACKLEY, A., HACHADOORIAN, R., ARENS, R., MAISLIN, G. and PACK, A. I. Identification of upper airway anatomic risk factors for obstructive sleep apnea with volumetric magnetic resonance imaging. *American journal of respiratory and critical care medicine*, vol. 168(5), pages 522–530, 2003. [2](#), [3](#), [12](#), [17](#), [103](#)
- SHEPERTYCKY, M. R., BANNO, K. and KRYGER, M. H. Differences between men and women in the clinical presentation of patients diagnosed with obstructive sleep apnea syndrome. *Sleep*, vol. 28(3), pages 309–314, 2005. [13](#)
- SHIVAKUMAR, P. G., LI, M., DHANDHANIA, V. and NARAYANAN, S. S. Simplified and supervised i-vector modeling for speaker age regression. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4833–4837. IEEE, IEEE, 2014. [73](#), [74](#)
- SMIALOWSKI, P., FRISHMAN, D. and KRAMER, S. Pitfalls of supervised feature selection. *Bioinformatics*, vol. 26(3), pages 440–443, 2010. ISSN 1460-2059. [5](#), [100](#), [102](#)
- SOLÀ-SOLER, J., FIZ, J. A., MORERA, J. and JANÉ, R. Multiclass classification of subjects with sleep apnoea–hypopnoea syndrome through snoring analysis. *Medical Engineering & Physics*, vol. 34(9), pages 1213–1220, 2012. ISSN 13504533. [3](#)
- SOLÀ-SOLER, J., FIZ, J. A., TORRES, A. and JANE, R. Identification of Obstructive Sleep Apnea patients from tracheal breath sound analysis during wakefulness in polysomnographic studies. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4232–4235. IEEE, IEEE, 2014. ISBN 978-1-4244-7929-0. [91](#)
- SOLÀ-SOLER, J., JANE, R., FIZ, J. A. and MORERA, J. Formant frequencies of normal breath sounds of snorers may indicate the risk of Obstructive Sleep Apnea Syndrome. In *2008 30th*

- Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3500–3503. IEEE, IEEE, 2008. ISBN 978-1-4244-1814-5. [3](#)
- SOLÉ-CASALS, J., MUNTEANU, C., MARTÍN, O. C., BARBÉ, F., QUEIPO, C., AMILIBIA, J. and DURÁN-CANTOLLA, J. Detection of severe obstructive sleep apnea through voice analysis. *Applied Soft Computing*, vol. 23, pages 346–354, 2014. ISSN 15684946. [5](#), [44](#), [62](#), [68](#), [86](#), [87](#), [88](#), [89](#), [91](#), [96](#), [99](#), [100](#), [144](#), [147](#), [158](#)
- SOONG, F. K., ROSENBERG, A. E., JUANG, B.-H. and RABINER, L. R. Report: A Vector Quantization Approach to Speaker Recognition. *AT&T Technical Journal*, vol. 66(2), pages 14–26, 1987. ISSN 87562324. [34](#)
- STRADLING, J. R. and CROSBY, J. H. Predictors and prevalence of obstructive sleep apnoea and snoring in 1001 middle aged men. *Thorax*, vol. 46(2), pages 85–90, 1991. [9](#), [103](#)
- STROHL, K. P. and REDLINE, S. Recognition of obstructive sleep apnea. *American journal of respiratory and critical care medicine*, vol. 154(2), pages 279–289, 1996. [12](#), [14](#)
- SUN, Y., WANG, X. and TANG, X. Deep Convolutional Network Cascade for Facial Point Detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483. IEEE, 2013. ISBN 978-0-7695-4989-7. [107](#), [125](#)
- SUN, Y., WANG, X. and TANG, X. Deep Learning Face Representation from Predicting 10,000 Classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898. IEEE, 2014. ISBN 978-1-4799-5118-5. [125](#), [128](#)
- TAIGMAN, Y., YANG, M., RANZATO, M. and WOLF, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708. IEEE, 2014. ISBN 978-1-4799-5118-5. [128](#)
- TITZE, I. R. Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, vol. 85(4), pages 1699–1707, 1989. ISSN 0001-4966. [100](#)
- TODD, E. S., WEINBERG, S. M., BERRY-KRAVIS, E. M., SILVESTRI, J. M., KENNY, A. S., RAND, C. M., ZHOU, L., MAHER, B. S., MARAZITA, M. L. and WEESE-MAYER, D. E. Facial Phenotype in Children and Young Adults with PHOX2B-Determined Congenital Central Hypoventilation Syndrome: Quantitative Pattern of Dysmorphology. *Pediatric Research*, vol. 59(1), pages 39–45, 2006. ISSN 0031-3998. [2](#)

- TOLEDANO, D., HERNANDEZ GOMEZ, L. and GRANDE, L. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, vol. 11(6), pages 617–625, 2003. ISSN 1063-6676. [36](#)
- TSANAS, A., LITTLE, M. A., MCSHARRY, P. E. and RAMIG, L. O. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of the Royal Society Interface*, vol. 8(59), pages 842–855, 2011. [2](#)
- TURK, M. and PENTLAND, A. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591. IEEE Comput. Soc. Press, Maui, HI, USA, USA, 1991. ISBN 0-8186-2148-6. [106](#)
- TYAN, M., ESPINOZA-CUADROS, F., FERNÁNDEZ POZO, R., TOLEDANO, D., LOPEZ GONZALO, E., ALCAZAR RAMIREZ, J. D. and HERNANDEZ GOMEZ, L. A. Obstructive Sleep Apnea in Women: Study of Speech and Craniofacial Characteristics. *JMIR mHealth and uHealth*, vol. 5(11), page e169, 2017. ISSN 2291-5222. [31](#), [147](#), [148](#)
- VALIPOUR, A. Gender-related Differences in the Obstructive Sleep Apnea Syndrome. *Pneumologie*, vol. 66(10), pages 584–588, 2012. ISSN 0934-8387. [141](#)
- VALIPOUR, A., LOTHALLER, H., RAUSCHER, H., ZWICK, H., BURGHUBER, O. C. and LAVIE, P. Gender-Related Differences in Symptoms of Patients With Suspected Breathing Disorders in Sleep: A Clinical Population Study Using the Sleep Disorders Questionnaire. *Sleep*, vol. 30(3), pages 312–319, 2007. ISSN 1550-9109. [139](#)
- VAPNIK, V. *The nature of statistical learning theory*. Springer science & business media, 2013. [42](#), [43](#)
- VEROPOULOS, K., CAMPBELL, C. and CRISTIANINI, N. Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on AI*, pages 55–60. 1999. [42](#)
- VGONTZAS, A. N., BIXLER, E. O., TAN, T.-L., KANTNER, D., MARTIN, L. F. and KALES, A. Obesity without sleep apnea is associated with daytime sleepiness. *Archives of internal medicine*, vol. 158(12), pages 1333–1337, 1998. [10](#)
- VINER, S., SZALAI, J. P. and HOFFSTEIN, V. Are history and physical examination a good screening test for sleep apnea? *Annals of internal medicine*, vol. 115(5), pages 356–359, 1991. [15](#)

- WALDRON, I. What do we know about causes of sex differences in mortality? A review of the literature. *Population Bulletin of the United Nations*, (18), pages 59–76, 1985. [13](#)
- WANG, Y. and KOSINSKI, M. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images., 2017. [128](#), [130](#), [134](#)
- WHITE, D. P., GIBB, T. J., WALL, J. M. and WESTBROOK, P. R. Assessment of accuracy and analysis time of a novel device to monitor sleep and breathing in the home. *Sleep*, vol. 18(2), pages 115–126, 1995. [54](#), [55](#)
- WHITTLE, A. T., MARSHALL, I., MORTIMORE, I. L., WRAITH, P. K., SELLAR, R. J. and DOUGLAS, N. J. Neck soft tissue and fat distribution: comparison between normal men and women by magnetic resonance imaging. *Thorax*, vol. 54(4), pages 323–8, 1999. ISSN 0040-6376. [140](#), [150](#), [151](#)
- WILL, M. J., ESTER, M. S., RAMIREZ, S. G., TINER, B. D., MCANEAR, J. T. and EPSTEIN, L. Comparison of cephalometric analysis with ethnicity in obstructive sleep apnea syndrome. *Sleep*, vol. 18(10), pages 873–5, 1995. ISSN 0161-8105. [13](#)
- WILTSHIRE, N., KENDRICK, A. H. and CATTERALL, J. R. Home oximetry studies for diagnosis of sleep apnea/hypopnea syndrome: limitation of memory storage capabilities. *CHEST Journal*, vol. 120(2), pages 384–389, 2001. [16](#)
- WIMMS, A., WOHRLE, H., KETHEESWARAN, S., RAMANAN, D. and ARMITSTEAD, J. Obstructive Sleep Apnea in Women: Specific Issues and Interventions. *BioMed Research International*, vol. 2016, pages 1–9, 2016. ISSN 2314-6133. [139](#), [140](#)
- WISKOTT, L., FELLOUS, J.-M., KUIGER, N. and VON DER MALSBERG, C. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(7), pages 775–779, 1997. ISSN 0162-8828. [106](#)
- YOUNG, T. and FINN, L. Epidemiological insights into the public health burden of sleep disordered breathing: sex differences in survival among sleep clinic patients. *Thorax*, vol. 53 Suppl 3(suppl 3), pages 16–9, 1998. ISSN 0040-6376. [140](#)
- YOUNG, T., FINN, L., AUSTIN, D. and PETERSON, A. Menopausal Status and Sleep-disordered Breathing in the Wisconsin Sleep Cohort Study. *American Journal of Respiratory and Critical Care Medicine*, vol. 167(9), pages 1181–1185, 2003. ISSN 1073-449X. [140](#)

- YOUNG, T., HUTTON, R., FINN, L., BADR, S. and PALTA, M. The gender bias in sleep apnea diagnosis: are women missed because they have different symptoms? *Archives of internal medicine*, vol. 156(21), pages 2445–2451, 1996. [13](#), [14](#), [140](#)
- YOUNG, T., PALTA, M., DEMPSEY, J., SKATRUD, J., WEBER, S. and BADR, S. The Occurrence of Sleep-Disordered Breathing among Middle-Aged Adults. *New England Journal of Medicine*, vol. 328(17), pages 1230–1235, 1993. ISSN 0028-4793. [9](#)
- YOUNG, T., PEPPARD, P. E. and GOTTLIEB, D. J. Epidemiology of obstructive sleep apnea: a population health perspective. *American journal of respiratory and critical care medicine*, vol. 165(9), pages 1217–39, 2002a. ISSN 1073-449X. [13](#)
- YOUNG, T., PEPPARD, P. E. and TAHERI, S. Excess weight and sleep-disordered breathing. *Journal of applied physiology (Bethesda, Md. : 1985)*, vol. 99(4), pages 1592–9, 2005. ISSN 8750-7587. [103](#)
- YOUNG, T., SHAHAR, E., NIETO, F. J., REDLINE, S., NEWMAN, A. B., GOTTLIEB, D. J., WALSLEBEN, J. A., FINN, L., ENRIGHT, P., SAMET, J. M. and SLEEP HEART HEALTH STUDY RESEARCH GROUP. Predictors of sleep-disordered breathing in community-dwelling adults: the Sleep Heart Health Study. *Archives of internal medicine*, vol. 162(8), pages 893–900, 2002b. ISSN 0003-9926. [11](#)
- YUILLE, A. L., HALLINAN, P. W. and COHEN, D. S. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, vol. 8(2), pages 99–111, 1992. ISSN 0920-5691. [107](#)
- ZHAO, W., CHELLAPPA, R., PHILLIPS, P. J. and ROSENFELD, A. Face recognition: A literature survey. *ACM Computing Surveys*, vol. 35(4), pages 399–458, 2003. ISSN 03600300. [105](#)
- ZIGEL, Y., TARASIUK, A. and GOLDSSTEIN, E. Analysis of speech signals among obstructive sleep apnea patients. In *Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of*, pages 760–764. IEEE, 2008. [2](#), [3](#), [5](#), [24](#), [63](#)