TELECOMUNICACIÓN

Campus Sur

POLITÉCNICA

PROYECTO FIN DE GRADO

TÍTULO: Análisis de métodos de parametrización y clasificación para la simulación de un sistema de evaluación perceptual del grado de afección en voces patológicas

AUTOR: Laureano Moro Velázquez

TUTOR (o Director en su caso): Juan Ignacio Godino Llorente

DEPARTAMENTO: Ingeniería de Circuitos y Sistemas - ICS

TITULACIÓN: Sonido e Imagen

VºBº

Miembros del Tribunal Calificador:

PRESIDENTE: Luis Narvarte Fernández

TUTOR: Juan Ignacio Godino Llorente

SECRETARIO: Juan Carlos González de Sande

Fecha de lectura: **de** **de**

Calificación:

El Secretario,

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA UNIVERSITARIA DE INGENIERÍA TÉCNICA DE TELECOMUNICACIÓN

PROYECTO FIN DE GRADO

Análisis de métodos de parametrización y clasificación para la simulación de un sistema de evaluación perceptual del grado de afección en voces patológicas

Autor: Laureano Moro Velázquez
Tutor: Juan Ignacio Godino Llorente

Junio de 2013

Resumen

Los procedimientos de evaluación de la calidad de la voz basados en la valoración subjetiva a través de la percepción acústica por parte de un experto están bastante extendidos. Entre ellos, el protocolo GRBAS es el más comúnmente utilizado en la rutina clínica. Sin embargo existen varios problemas derivados de este tipo de estimaciones, el primero de los cuales es que se precisa de profesionales debidamente entrenados para su realización. Otro inconveniente reside en el hecho de que, al tratarse de una valoración subjetiva, múltiples circunstancias significativas influyen en la decisión final del evaluador, existiendo en muchos casos una variabilidad inter-evaluador e intra-evaluador en los juicios. Por estas razones se hace necesario el uso de parámetros objetivos que permitan realizar una valoración de la calidad de la voz y la detección de diversas patologías.

Este trabajo tiene como objetivo comparar la efectividad de diversas técnicas de cálculo de parámetros representativos de la voz para su uso en la clasificación automática de escalas perceptuales. Algunos parámetros analizados serán los coeficientes *Mel-Frequency Cepstral Coefficients* (MFCC), las medidas de complejidad y las de ruido. Así mismo se introducirá un nuevo conjunto de características extraídas del Espectro de Modulación (EM) denominadas Centroides del Espectro de Modulación (CEM). En concreto se analizará el proceso de detección automática de dos de los cinco rasgos que componen la escala GRBAS: G y R. A lo largo de este documento se muestra cómo las características CEM proporcionan resultados similares a los de otras técnicas anteriormente utilizadas y propician en algún caso un incremento en la efectividad de la clasificación cuando son combinados con otros parámetros.

Agradecimientos

Quiero agradecer enormemente a los integrantes del grupo “Bioingeniería y Optoelectrónica – BYO” del departamento ICS su constante ayuda. Vuestra contagiosa pasión por la investigación me inspira día a día al tiempo que me enriquece personal y profesionalmente.

Este trabajo ha sido financiado por el proyecto TEC2012-38630-C04-01 "EVALUACION MULTIMODAL DE TRASTORNOS NEUROLOGICOS MEDIANTE LA CARACTERIZACION DE LA VOZ, DINAMICA DE LOS PLIEGUES VOCALES Y SECUENCIAS SACADICAS" del Ministerio de Economía y Competitividad del Gobierno de España.

Índice de contenidos

Índice de tablas

Índice de figuras

Glosario de acrónimos y abreviaturas

A

A	Aesthenia
AD	Ambiguity Domain
AE	Aproximate Entropy
ANN	Artificial Neural Network
ATRI	Amplitude Tremor Intensity Index
AAW	Acoustic Average Wave

B

B	Breathiness

C

CAPE-V	Consensus Auditory Perceptual Evaluation of Voice
CD	Correlation Dimension
CEM	Centroides del Espectro de Modulación
CHNR	Cepstrum Harmonic to noise Ratio
COG	Center Of Gravity

D

DFA	Detrended Fluctuation Analysis
DFT	Discrete Fourier Transform

E

EGG	Electroglotografía
EM	Espectro de Modulación
EMD	Empirical Mode Decomposition

F

FE	Fuzzy Entropy
FTRI	F_0 Tremor Intensity Index

G

G	Grade
GKAE	Gaussian Kernel Aproximate Entropy
GMM	Gausian Mixture Models
GNE	Glottal to Noise Excitation ratio

H

HMM	Hidden Markov Models
K-NN	K-nearest neighbours
HNR	Harmonic to Noise Ratio
HOS	Higher Order Statistic
HOSVD	Higher Order Singular Value Decomposition

L

LDA	Linear Discriminant Analysis
LFSC	Linear Frequency Spectrum Coefficients
LPC	Linear Predictive Coding
LLE	Largest Lyapunov Exponent
LVQ	Learning Vector Quantifiers

M

MDVP	Multidimensional Voice Program
MEEI	Massachusetts Eye and Ear Infirmary
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multilayer Perceptrons
MRFSC	Maximal Relevance Feature Selection Criterion
MSE	Modified Sampled Entropy
MWC	Mucosal Wave Correlate

N

NNE	Normalized Noise Energy

P

PLI	Pathological Likelihood Index
PSD	Power Spectral Density

R

R	Roughness
RBH	Roughness, Breathiness and Hoarseness
RPDE:	Recurrence Period Density Entropy

S

S	Strain
SE	Sampled Entropy
SOM	Self Organizing Maps
SPI	Soft Phonation Index
STFT	Short-Time Fourier Transform
SVM	Support Vector Machines

T

TFM	Time-Frequency Matrix

U

UBM	Universal Background Models

V

VTI	Voice Turbulence Index

1. Introducción

En este capítulo se introducirán las distintas técnicas de evaluación de la calidad de la voz y diagnóstico de enfermedades relacionadas con el aparato fonador. Posteriormente se expondrán los puntos básicos del protocolo de evaluación perceptual GRBAS [1] ya que se hará referencia a él a lo largo del proyecto.

1.1. Evaluación de las patologías de la voz

El ser humano utiliza su voz como herramienta de comunicación con su entorno, expresando así ideas o sentimientos. Debido a muy diversas causas, el aparato fonador de una persona puede no funcionar correctamente o estar dañado, lo que en algunos casos interferiría en un proceso de comunicación normal. Por estos motivos se hace necesaria la evaluación de las patologías de la voz al igual que su diagnóstico, para así poder aplicar un determinado tratamiento.

Según [2], los métodos de evaluación de la presencia de una patología en el aparato fonador pueden dividirse en varios grupos:

- Examen físico
- Examen visual y técnicas de imagen médica
- Estudio de la aerodinámica
- Análisis acústico de la voz
- Análisis perceptual de la calidad de la voz

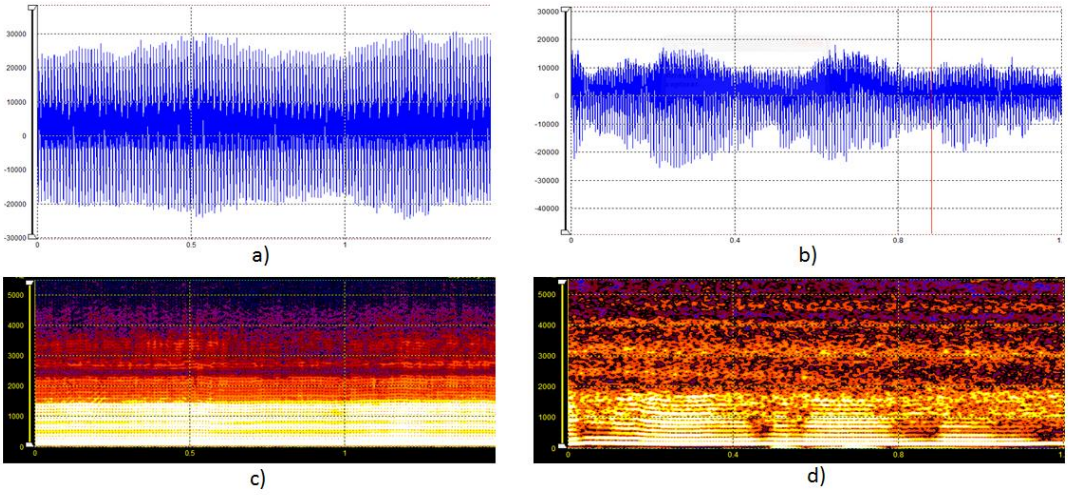
El **examen físico** consiste en la observación completa de la fisonomía de la cabeza y el cuello mediante inspección visual y palpación, realizándose a la vez pruebas de la función motora de las distintas estructuras. De esta forma se busca la existencia de distintos problemas que puedan dificultar la generación de voz.

El **examen visual** suele estar relacionado con las técnicas de observación de la morfología y funcionamiento de la laringe. Normalmente este examen es realizado con un espejo laríngeo, aunque si se desea profundizar deberán utilizarse **técnicas laringoscópicas**, laringoestroboscópicas o grabaciones de vídeo de alta velocidad. Estas

últimas permitirán conocer el funcionamiento y el aspecto de las cuerdas vocales con mayor detalle.

Durante el proceso de producción de la voz, la presión, el flujo y la resistencia del aire expulsado están interrelacionados y aportan información sobre el modelo de generación de voz y sus posibles deficiencias. A la observación de este tipo de señales se le denomina **estudio de la aerodinámica**.

El **análisis acústico** de la voz resulta ideal en la monitorización de la evolución de un paciente tras diversos tratamientos. Esto es debido a la sencillez de obtención de la señal y los recursos necesarios, generalmente asequibles desde puntos de vista económicos y de usabilidad. A estos factores debemos sumarles la efectividad de muchas de las recientes técnicas de procesado de señales. Estas suelen proporcionar diversos parámetros numéricos obtenidos a partir de la información espectral y temporal de la voz, los cuales están relacionados con la presencia de algunas patologías. En la Figura 1 puede observarse la forma de onda y el espectrograma de dos señales acústicas: una voz normal y una patológica. En ella es fácil ver las



diferencias tanto en el dominio del tiempo como en el de la frecuencia.

Figura 1. Formas de onda y espectrogramas de una voz normal (a y c) y de una patológica (b y d) afectada por un quiste epidermoide. Imágenes obtenidas con el software WPCVox.

Por último, el **análisis perceptual** consiste en una evaluación de determinadas características de la voz por parte de un experto (foniatra u otorrinolaringólogo, por ejemplo). Se basa en la escucha de una vocal sostenida o de un fragmento de texto

leído por el paciente, tras lo cual se suele dictaminar un valor numérico relacionado con el grado de perturbación existente en dicha voz. Una de las escalas más utilizada y extendida es la escala GRBAS.

Otras escalas de evaluación son la *Buffalo Rating Voice Profile* [3], que toma valores de 1 a 5 en función del grado y extensión de las perturbaciones de la voz percibidas; el procedimiento *Roughness, Breathiness and Hoarseness* (RBH) [4] en el que se evalúan tres índices que varían entre 0 y 3; o el más reciente protocolo *Consensus Auditory Perceptual Evaluation of Voice* (CAPE-V) [5], desarrollado por la institución denominada *American Speech Language Hearing Association*.

El problema del análisis perceptual de la calidad de la voz es su subjetividad y los factores externos que lo influencian, como el estado de ánimo del evaluador, su entrenamiento previo, cansancio, estrés, su cultura, etc [6] [7]. Por eso se hace necesario buscar un sistema automático de clasificación que sea capaz de emular el análisis perceptual de un experto pero de una manera más objetiva y reproducible, reduciendo de este modo la incertidumbre de las valoraciones realizadas.

1.2. GRBAS

Este proyecto utiliza para el análisis propuesto la escala perceptual denominada GRBAS, presentada por el grupo de investigación japonés del Comité para Tests de Función Fonatoria de la Sociedad Japonesa de Logopedia y Foniatría en el año 1969. En gran medida se suelen seguir las indicaciones dadas por Hirano en su libro de 1981 [1], en el que se describen los pasos para realizar una valoración perceptual GRBAS. Según [8] el protocolo GRBAS es el más fiable de todos los métodos de evaluación de su clase en cuanto a variabilidad intra-evaluador e inter-evaluador, lo que lo convierte en un criterio de interés para los objetivos de nuestro trabajo.

La escala GRBAS está subdividida en cinco rasgos:

- G: “Grade”, grado general de disfonía. Representa una valoración global de la calidad de la voz. Se suele asociar el término disfonía a la existencia de cualquier tipo de desviación patológica percibida en la voz. Estas desviaciones pueden estar relacionadas con el nivel, el pitch, el timbre o incluso la prosodia.
- R: “Roughness”, indica cómo de ruidosa e irregular es una voz y cómo de irregulares son los pulsos glóticos desde un punto de vista perceptual. Con este rasgo también se valora la existencia de tonos diferentes simultáneos durante la fonación (diplofonía) y las “roturas de la voz”. Se asocia normalmente con

lesiones orgánicas que producen una disminución de la vibración de las cuerdas vocales

- B: “Breathiness”, indica cuánto flujo de aire adicional se percibe en la voz del paciente. Suele estar relacionado con lesiones orgánicas que producen fallos en el cierre de las cuerdas vocales, lo que provoca que una cierta cantidad de aire se pierda y no se utilice en la generación de la señal fundamental. Este tipo de voces llevan asociado un cierto ruido y una disminución de la intensidad. El rasgo B aumenta en casos de extrema rigidez de la mucosa en ausencia de hendidura glótica.

- A: “Aesthenia”, está referido al grado de debilidad de la voz. Este rasgo será más alto cuanto menor sea la intensidad producida. Está relacionado con una reducción de la función de los músculos de la laringe, lo que suele provocar fatiga vocal. Un ejemplo de enfermedad que produce un aumento en el nivel del rasgo A es la patología neurológica denominada miastenia gravis.

- S: “Strain”, es un indicativo de la hiperfunción de la voz relacionada con el aumento de la actividad muscular extrínseca de la laringe, provocando un sonido forzado. Es opuesto al rasgo A. Se da en casos de disfonía espasmódica de aducción entre otras patologías.

Cada uno de ellos podrá tener un nivel de 0 a 3, en función del grado de afección:

- 0: Normal
- 1: Leve
- 2: Moderado
- 3: Intenso

La evaluación de la escala GRBAS se realiza a través de la escucha de una vocal sostenida y de un determinado pasaje de texto. La notación habitual suele ser del tipo: $G_gR_rB_bA_aS_s$, siendo los subíndices g, r, b, a y s los niveles asociados a cada rango.

2. Mecanismos de fonación y patologías relacionadas

En este capítulo se describirá someramente el funcionamiento del aparato fonador así como las principales partes que lo componen. Posteriormente se introducirán los tipos de patologías de la voz más relevantes, su etiología y los síntomas que suelen presentar.

2.1. Anatomía y fisiología del aparato fonador

En este apartado se muestra la estructura y el funcionamiento del aparato fonador, en el cual cabe destacar tres subsistemas componentes: el aparato respiratorio, la laringe y las distintas cavidades superiores y resonadores. La Figura 2 muestra un esquema general de las diferentes partes del aparato fonador.

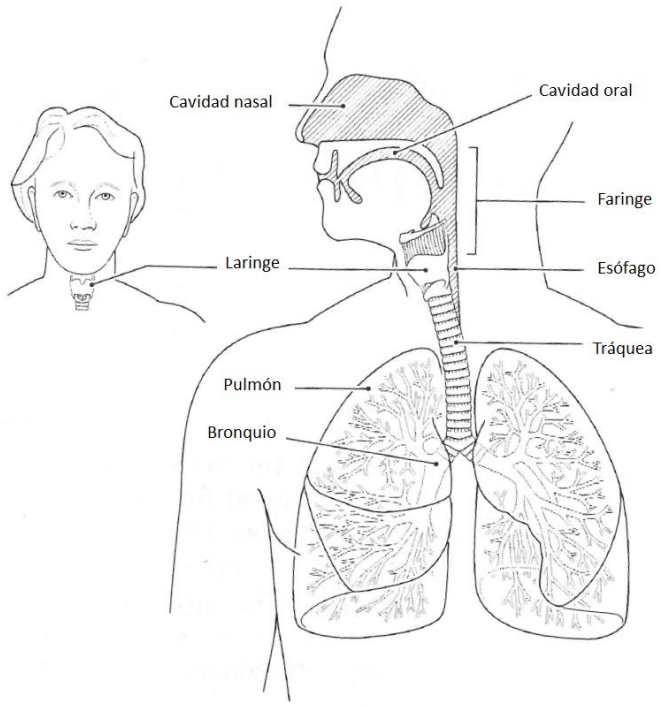


Figura 2. Esquema general del aparato fonador. Adaptado de [9].

A modo de resumen se puede decir que el aparato respiratorio es el encargado de inyectar un determinado flujo de aire a través de la laringe, donde se encuentran las cuerdas vocales, cuya vibración al paso del aire genera una señal básica. Esta es modificada en su camino por la laringe y las cavidades nasal y bucal y es radiada por distintas partes del cuerpo como la boca, frente, los pómulos o el pecho.

2.1.1. Aparato respiratorio

El aparato respiratorio no sólo tiene la misión de ventilación para el intercambio de O_2 y CO_2 necesario para mantener las funciones vitales, sino que juega un papel fundamental en la generación de la señal de voz. Los elementos básicos de este sistema son los pulmones, la caja torácica, la tráquea y el diafragma.

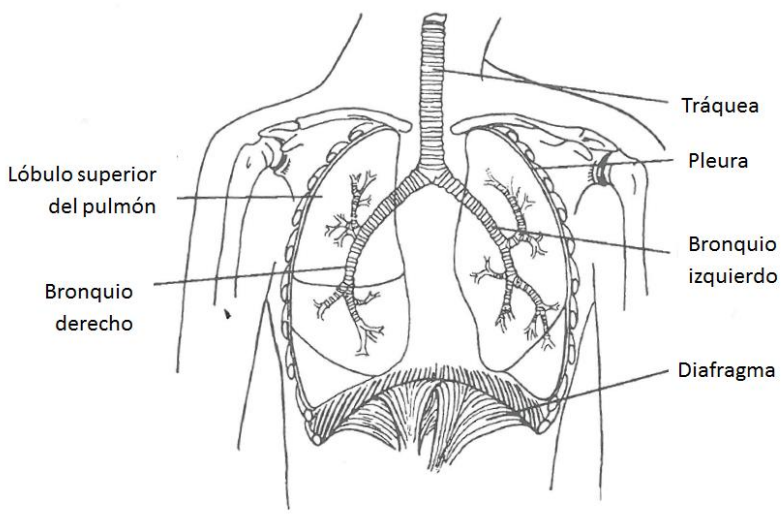


Figura 3. Esquema del aparato respiratorio adaptado de [2]

Los pulmones, formados por tejido elástico, se inflan y desinflan para propiciar el movimiento de aire. En ellos se encuentran los bronquios, como puede observarse en la Figura 3. Es en los bronquios donde se produce el intercambio de gases con los capilares sanguíneos. La tráquea permite la circulación de aire entre los pulmones y el exterior. La caja torácica protege los pulmones e interviene en el proceso respiratorio por acción del diafragma, el mayor músculo del aparato respiratorio que además separa el pecho del abdomen.

2.1.2. Laringe

La laringe se sitúa debajo de la faringe y encima de la tráquea, si bien su posición varía al tragar para que los líquidos o comidas puedan ser desviados de la faringe al esófago. La mayor parte de su estructura está formada por cartílagos, membranas y pequeños músculos, de los cuales podemos distinguir entre dos tipos: los intrínsecos y los extrínsecos. Los primeros interconectan los cartílagos de la laringe entre sí y los segundos unen esta estructura a otras adyacentes. Las cuerdas vocales se encuentran en la mitad de la laringe y son parte fundamental de esta. Por encima de ellas se encuentra el ventrículo laríngeo, las cuerdas ventriculares o falsas cuerdas, la membrana cuadrangular y los repliegues ari-epiglóticos. Todos estos pliegues, especialmente las cuerdas vocales, consiguen sellar correctamente la laringe para evitar el paso de aire cuando es necesario.

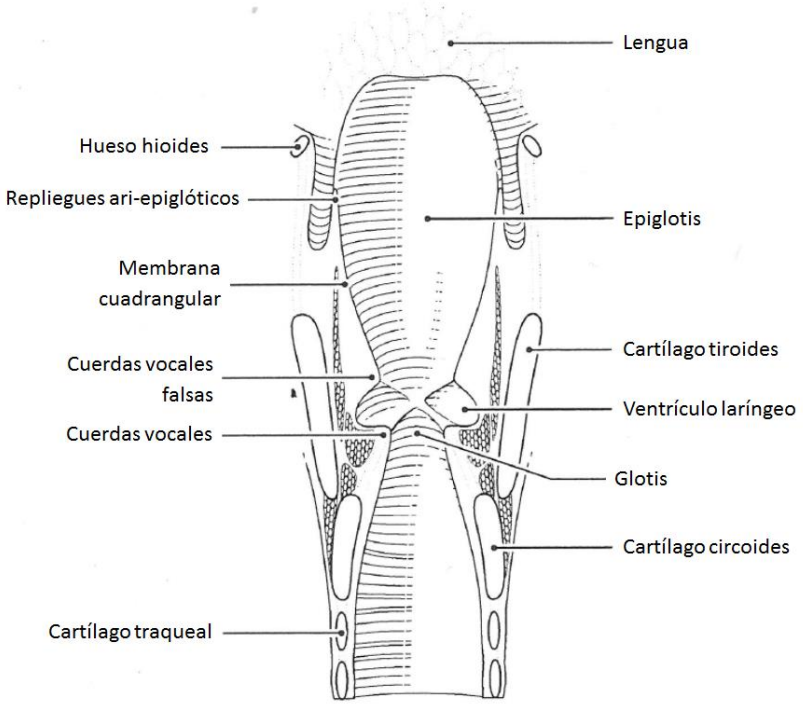


Figura 4. Partes principales de la laringe. Adaptado de [9]

Las cuerdas vocales están formadas por un epitelio en el exterior, seguido de otras tres capas de tejido que recubren el músculo tiroaritenoideo. Cuando están separadas aparece un espacio entre ellas llamado glotis, como se muestra en la Figura 4. A través de él circula el flujo de aire procedente de los pulmones produciéndose la fonación. Básicamente, el aire existente bajo la glotis cuando se encuentra cerrada,

está a una mayor presión que el aire existente en la parte superior. Esto ocurre principalmente por la presión que ejerce el diafragma sobre el volumen de aire contenido en los pulmones. Esa diferencia de presiones hace que las cuerdas vocales se desplacen hacia arriba y el flujo de aire pase por la glotis. Una vez se libera parte del aire, se genera un gradiente de presión entre la parte superior y la inferior de las cuerdas vocales que es contrario al existente antes de producirse la salida de aire. De este modo las cuerdas vuelven a su posición inicial volviéndose a cerrar la glotis. El proceso se repite con una frecuencia que normalmente puede ir de los 50 a los 300 Hz y da lugar a una onda de presión sonora de apariencia cuasi sinusoidal.

2.1.3. Tracto vocal: Laringe y cavidad bucal

La onda de presión sonora generada tras el paso del aire por la glotis sigue su camino por diversas cavidades, que la transforman debido a múltiples reflexiones y a la creación de ondas estacionarias. Una reflexión puede darse cuando una onda choca contra alguna de las paredes del tracto o cuando se produce un cambio de impedancias (cambio de sección). La laringe y la cavidad bucal suponen varios cambios de sección en el camino de la onda de presión hacia el exterior, según se muestra en la Figura 5. Normalmente las ondas estacionarias están provocadas por la suma de una onda incidente y una reflejada. El paso de la onda por estas cavidades conlleva diversas reflexiones y ondas estacionarias momentáneas, modelándose la señal que salía originalmente de la glotis dando lugar a la señal de la voz.

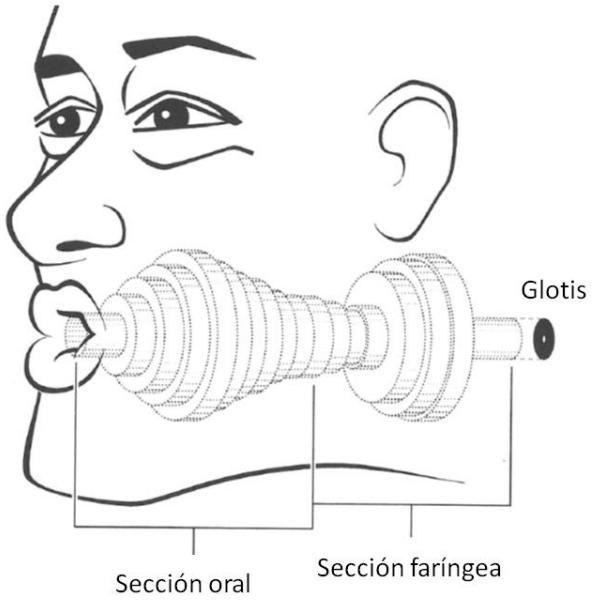


Figura 5. Cavidades superiores. Adaptado de [9]

De este modo, la señal generada en la glotis, que se caracteriza por una componente de frecuencia fundamental y un buen número de armónicos, es filtrada por el tracto vocal de acuerdo a una función de transferencia que determina los formantes, que son aquellas bandas de frecuencia en las que existen resonancias apreciables. En la Figura 6 se observa la forma de onda de los pulsos glotales y la producida por varios formantes.

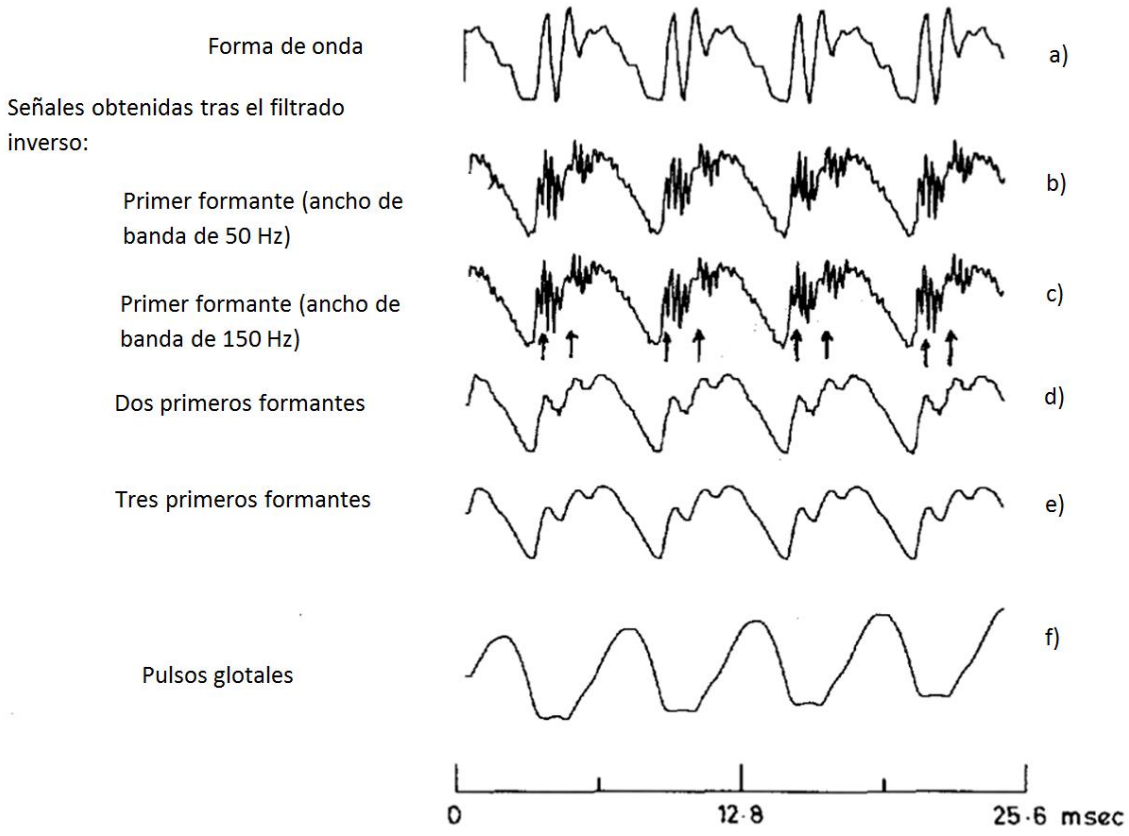


Figura 6. Forma de onda de la vocal /a/(a), forma de onda de la señal filtrada conteniendo varios formantes (b, c, d y e) y pulsos glotales (f). Adaptada de [10].

En función de la configuración del tracto vocal, se producirán unos formantes u otros. Básicamente, son los dos primeros los que determinan a qué vocal pertenece un segmento acústico. En la Figura 7 se representa la relación entre los dos primeros formantes y la vocal producida.

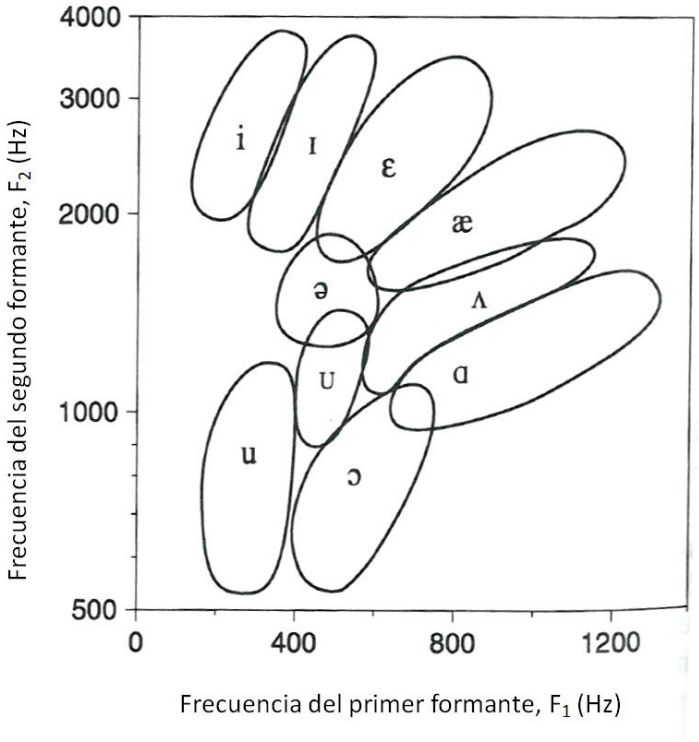


Figura 7. Representación de la relación frecuencial entre los dos primeros formantes para cada una de las vocales. Adaptado de [9].

2.2. Patologías de la voz

Es posible encontrar diversas patologías que afectan a la generación de voz, existiendo variedad de causas, técnicas de diagnóstico y tratamientos. Según [2], en función de la etiología, estas enfermedades se pueden clasificar como:

- Estructurales (cambios en las cuerdas vocales): Nódulos, pólipos, edema de Reinke, laringitis, granuloma, quistes, papiloma de laringe, tumor granular, sulcus, hiperqueratosis, cáncer de laringe o trauma de laringe entre otras.

- Neurológicas (por alteraciones en el sistema nervioso central o el periférico): Parálisis lateral de las cuerdas vocales, parálisis bilateral de las cuerdas vocales, parálisis del nervio laríngeo superior, disfonía espasmódica, tremor esencial, hipofonía asociada a la enfermedad de Parkinson, esclerosis lateral amiotrófica y atrofia múltiple entre otras.

- De lesión vascular: Varices, ecstasia y hemorragia de las cuerdas vocales.

- Funcional: Puberfonía, fonación ventricular, disfonía y afonía.

- Idiopático (cuando hay causas desconocidas para la enfermedad): Movimiento paradójico de las cuerdas vocales.

Las **patologías estructurales** provienen en muchos casos de anormalidades congénitas en las zonas glotales, subglotales y supraglotales. En otras ocasiones se deben al excesivo estrés o las prolongadas colisiones de las cuerdas vocales entre sí, dando lugar a nódulos. Si los nódulos no son tratados en una etapa temprana, irán haciéndose cada vez más duros y rígidos siendo necesaria la cirugía como tratamiento. Algo similar ocurre con los pólipos, que aparecen a causa del estrés mecánico añadido a algún tipo de irritación de los tejidos. Otras dolencias como la faringitis implican inflamaciones, en muchos casos crónicas, debidas a irritaciones provocadas en numerosas ocasiones por agentes como el tabaco.

Cuando las cuerdas vocales presentan una apariencia normal pero su vibración o funcionamiento no lo es, probablemente la etiología de la afección será de tipo **neurológico**. Uno de los casos más comunes de este tipo es la parálisis de una o de las dos cuerdas vocales, lo que también conlleva problemas respiratorios, al no poder realizar la laringe su función de válvula. La enfermedad de Parkinson o la esclerosis lateral amiotrófica suelen producir debilidad muscular, lo que provoca una voz débil e inestable. Un síntoma típico es el tremor o temblor de la voz en la que la oscilación incontrolada de los músculos asociados a la fonación produce un temblor perceptible de la voz.

Los trastornos relacionados con el estrés mecánico de la laringe serían las hemorragias o las varices, que se dan en casos de gritos continuados y de elevado nivel. Se trata de patologías de **lesión vascular**.

Un ejemplo de patología **funcional** es la puberfonía, que tiene lugar cuando la laringe no crece de manera adecuada tras la pubertad y el individuo mantiene una voz aguda. Por otro lado, podemos encontrar la disfonía, que aparece en casos de estrés o ansiedad. La disfonía está provocada por un incremento de la actividad muscular dando lugar a una voz tensa y aguda que a veces se convierte en débil y con exceso de ruido de aire.

Por último, como patología **idiopática** encontramos el movimiento paradójico de las cuerdas vocales que ocurre cuando se da la aducción de las cuerdas durante la inspiración. De este modo se provoca un sonido agudo asociado a la respiración, tos crónica, afonía o ronquera entre otros síntomas.

3. Estado del arte

En este capítulo se realiza una introducción a los distintos esquemas metodológicos que se suelen utilizar en sistemas de parametrización-clasificación de voces patológicas. Posteriormente se analizan los diferentes parámetros y clasificadores más comúnmente utilizados, y finalmente se describen algunos trabajos de investigación destacados o relevantes para el desarrollo de este mismo proyecto.

3.1. Esquemas metodológicos. Parametrización y clasificación

Como ya se explicó anteriormente, existen diversos métodos de detección de patologías de la voz mediante técnicas de imagen o electroglotografía (EGG), análisis aerodinámico o examen físico entre otras. Este trabajo se centra en el análisis acústico, que implica la adquisición de la señal de voz para realizar un procesado que permita la detección automática de patologías o el etiquetado de dicha voz dentro de una determinada escala. En este campo es posible encontrar múltiples estudios que se basan en el uso de parámetros y clasificadores con un determinado objetivo. En función de ese objetivo es posible identificar tres grandes grupos. El primer grupo se centra en discriminar entre una voz patológica y una normal. El segundo en distinguir automáticamente entre diversas patologías o categorías. Por ejemplo, permitiría discernir entre una voz que tuviese síntomas producidos por la presencia de nódulos y otra con parálisis de una de las cuerdas vocales. El tercer grupo se centra en proporcionar índices de calificación representativos de la calidad de la voz o simular escalas perceptuales de evaluación, como puede ser la escala GRBAS. En este tercer grupo se encuentra el presente estudio si bien se ha de prestar atención a los tres, debido a que las técnicas y metodologías que se utilizan pueden ser igualmente útiles independientemente del objetivo concreto de cada investigación.

El proceso habitual para la realización de estos sistemas de parametrización-clasificación se muestra en el diagrama de la Figura 8.

Para llegar a desarrollar un sistema como el del diagrama, es necesaria una fase de modelización en la cual se analizan las distintas características se pueden extraer, su relevancia, los tipos de clasificadores existentes y su conveniencia, etc. De dicha fase se obtienen unas conclusiones que permiten realizar el modelo óptimo de parametrización-clasificación.

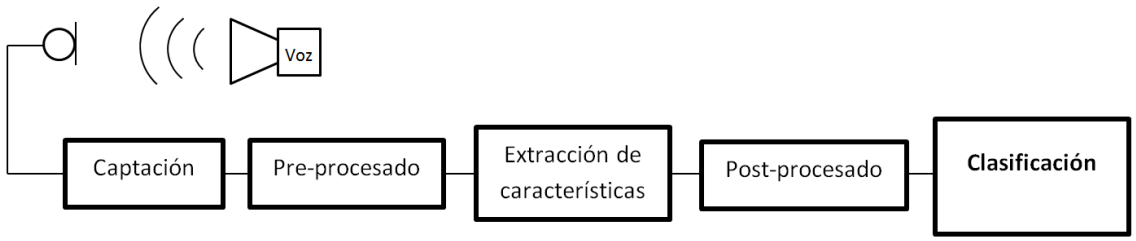


Figura 8. Diagrama tipo de un sistema de clasificación de grado de calidad de la voz.

El primer paso consiste en la captación de señales de distintas categorías incluyendo voces normales y con distintas patologías y grados de afección. Todas las muestras deben ser recogidas preferentemente en las mismas condiciones para evitar que la influencia del canal sesgue los resultados de clasificación. A este respecto, el uso de Universal Background Models (UBM) [11], [12] ayuda a independizar la clasificación de ciertos factores como el canal.

Una situación muy ventajosa desde el punto de vista científico se daría cuando la mayor parte de los estudios centrados en un área concreto usasen la misma base de datos. Esto permitiría la comparación de resultados de una manera más real y efectiva. Aunque esto no ocurre con todos los trabajos analizados, sí se puede decir que la base de datos más utilizada es la realizada por el *Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory*, comercializada por la empresa Kay Elemetrics [13]. Este conjunto de registros contiene grabaciones de 53 voces normales y 566 patológicas, incluyendo vocales sostenidas y fragmentos de habla continua.

Sin embargo, y ante la necesidad de uso de habla continua en otros idiomas o de análisis de patologías muy concretas, algunos estudios utilizan registros propios. Este es el caso del presente estudio, en el que se usará la base de datos del Hospital Príncipe de Asturias de Alcalá de Henares (PdA) que contiene grabaciones de 433 voces normales y 215 patológicas.

En la fase de preprocesado la señal suele normalizarse y segmentarse. En muchos casos, se hace necesaria la eliminación de fragmentos sin señal de voz, como se indica en [14]. La longitud de los segmentos puede ser muy variable dependiendo del parámetro a analizar. Algunos de ellos evalúan el comportamiento de la señal a nivel de periodo fundamental para poder extraer información sobre micro-variaciones. Otros son más globales y pueden llegar a necesitar de la señal completa para ser

representativos. Por ejemplo, en [15] se utilizan tramas de 262 ms para obtener información relevante del Espectro de Modulación (EM) mientras que en [16] las tramas tienen una duración de 40 ms, ya que se usan Mel Frequency Cepstral Coefficients (MFCC).

La parametrización se utiliza para extraer las características que representen a los distintos registros de audio y que permitan clasificarlos de la forma más inequívoca posible. La mayor parte de los trabajos usan parámetros de perturbación de amplitud, perturbación de frecuencia, ruido o complejidad.

Algunos ejemplos de estos cuatro grupos de parámetros son:

- Perturbación de amplitud:
    - Shimmer: variaciones de los niveles de pico de amplitud entre periodos adyacentes [17].
    - Amplitude Tremor Intensity Index (ATRI): mide las variaciones de la amplitud de la señal a largo plazo [18]. Este tipo de medida también podría ser clasificado como parámetro de tremor, al igual que el F_0 Tremor Intensity Index (FTRI).
- Perturbación de frecuencia:
    - Jitter: variación de la frecuencia fundamental entre ciclos adyacentes [17].
    - F_0 Tremor Intensity Index (FTRI): mide las variaciones de la frecuencia fundamental de la voz a largo plazo [18].
    - Estudio del MFCC: los MFCC son tal vez los parámetros más utilizados para tareas de reconocimiento y análisis de patologías relacionadas con la voz [19], [20], [21]. Representan una distribución de la energía en bandas situadas sobre la escala perceptual Mel [22]. Debido a que en este proyecto también se utilizan, serán desarrollados en detalle más adelante.
    - Estudio del Espectro de modulación (EM): proporciona información sobre el espectro acústico y las modulaciones sobre este [15]. En capítulos posteriores se detallan las propiedades del EM.

  - Técnicas basadas en Linear Predictive Coding (LPC): sirven para extraer información sobre los formantes. La amplitud y ancho de banda de estos puede ser utilizada para detectar patologías [23].

- Ruido:

  - Relación Armónico-Ruido (HNR): permite caracterizar la presencia de ruido en la voz para poder utilizarlo como parámetro identificativo de las voces patológicas [24].

  - Relación excitación glotal-ruido (Glottal to Noise Excitation ratio - GNE): permite medir la cantidad de excitación generada en las cuerdas vocales frente al ruido de turbulencia ocasionado en la laringe [19], [25].

  - Cepstrum Harmonic to noise Ratio (CHNR) y Normalized Noise Energy (NNE): el primero permite estimar la energía del ruido a partir de un análisis cepstrum y el segundo mide la relación entre la energía del ruido y la energía total del espectro [19].

- Complejidad:

  - Características de no linealidad temporales: En ocasiones se utilizan distintos parámetros relacionados con la modulación en amplitud o en frecuencia de los formantes a lo largo del tiempo, exponentes Lyapunov, correlación o entropía para caracterizar las voces patológicas [26].

Adicionalmente, existen algunos parámetros como el Mucosal Wave Correlate (MWC) y Acoustic Average Wave (AAW) [27] que representan características biomecánicas del tracto vocal.

También se dan estudios que se apoyan en el uso de parámetros combinados tiempo-frecuencia, como Time-Frequency Descomposition [28], o que directamente fusionan descriptores de varios grupos de los enumerados previamente.

En general se observa que la parametrización y detección pueden realizarse basándose en el análisis de una vocal sostenida (normalmente /a/) o del habla continua. En el primer caso, la segmentación es mucho más simple y reduce las posibilidades de disminución de tasa de precisión por fallos en la segmentación. Por otro lado, hay características de la señal de voz que sólo podrán observarse con habla continua. Por estos motivos la mayoría de los trabajos observados (véase la Tabla 1) utilizan vocales sostenidas pero no todos.

En muchos casos, las características que se obtienen contienen información que no es relevante y que en muchos casos podría entorpecer la fase de clasificación. Para ello es recomendable optimizar los datos obtenidos en la parametrización a través de técnicas de selección de características o de reducción de información como son la *Maximal Relevance Feature Selection Criterion (MRFSC)* o *Higher Order Singular Value Decomposition* (HOSVD) utilizados en estudios como [15] y [16]. Este filtrado podría considerarse un tipo de post-procesado, al igual que la fusión o la normalización de características.

Finalmente se realiza la clasificación a través de un modelo, el cual, recibe las características que llegan de la fase anterior y toma la decisión sobre la clase pertenecen. Las clases pueden representar un tipo de enfermedad, como sulcus o nódulos, la existencia o no de una patología, o un nivel numérico que represente el grado de afección que tiene un paciente. Normalmente se usa parte de la base de datos para generar el modelo predictivo. A este tipo de entrenamiento de modelo se le llama aprendizaje supervisado. Los clasificadores más comúnmente utilizados son:

- Hidden Markov Models (HMM) y Gausian Mixture Models (GMM). Se crean modelos probabilísticos a partir de los datos observables que permiten la clasificación de una manera robusta y computacionalmente eficiente. Son usados en estudios como [21], [29] o [30] entre muchos otros.

- Redes neuronales artificiales. Permiten una relación no lineal entre la entrada y la salida. Están basadas en la creación de varias capas de neuronas (unidades de procesamiento muy simple) conectadas entre sí que encaminan el "estímulo de entrada" a través de la red hacia una determinada salida que representa la clase. Existen varios tipos de estructuras, entre las que encontramos los Multilayer Perceptrons (MLP), Learning Vector Quantifiers (LVQ) o Self Organizing Maps (SOM). Podemos encontrar ejemplos de su uso en [19], [20] y [31].

- K-nearest neighbours (K-NN). Con esta técnica se modelan las funciones de densidad probabilidad de cada clase y posteriormente, para cada vector de entrada, se calcula la probabilidad de que pertenezca a cada una de las clases, seleccionándose la clase que tenga mayor probabilidad a la salida. Los estudios [23], [24] y [32] usan este clasificador.

- Support Vector Machines (SVM) [33] [34]. Este tipo de clasificadores, se basan en la creación de hiperplanos con áreas representativas de cada clase. En función de la situación de los vectores de entrada se selecciona una clase u otra

como clase detectada. Este es el método utilizado en el presente proyecto por lo que será desarrollado más adelante. Pueden encontrarse ejemplos de su uso en [15] o [35].

Como se comentó anteriormente, para poder obtener uno de estos modelos se debe realizar un entrenamiento con los registros de audio de la base de datos. Tras el entrenamiento se realiza la validación del mismo con el objetivo de conocer su efectividad. Existen múltiples técnicas de validación como bootstrap o la validación cruzada [36], que utilizan parte de los audios de la base de datos, normalmente no utilizados en el entrenamiento, para comprobar el correcto funcionamiento del modelo.

3.2. Algunos trabajos de investigación

En el punto anterior se enumeran los tres tipos de estudios habituales en el análisis de voces patológicas en función de la clasificación que realizan. Estos son:

- Aquellos que se centran en la discriminación entre voces normales y patológicas
- Los que buscan discernir entre varias patologías
- Estudios enfocados en la creación de un modelo que simule una evaluación perceptual automática.

En referencia a la discriminación automática entre voces patológicas y normales la mayoría de los trabajos han logrado muy buenos resultados llegando a tasas de precisión de hasta el 98,7% según se muestra en la Tabla 1.

En [15] se utilizan características extraídas del EM para discriminar entre voces normales y patológicas. En él se obtiene una precisión superior al 94%. El trabajo de investigación [16], con un objetivo similar utiliza parámetros obtenidos de análisis MFCC y de EM los cuales, tras una primera fase de clasificación, dan lugar a vectores que se combinan y pasan a un segundo esquema de decisión. En este caso se obtienen precisiones superiores al 95%. En estos ejemplos, la dimensionalidad del EM se reduce usando la descomposición HOSVD. Debido a los buenos resultados que proporciona, el EM será tenido en cuenta como una fuente de características más a utilizar en la fase de parametrización dentro de este proyecto.

Parámetro	**Clasificador**	**BBDD, tipo de voz, (#Normal + #Patológica)**	**Precisión (%)**
HNR [24]	3-NN	MEEI, vocal sostenida (53 + 163)	94.3
Características temporales y espectrales clásicas [21]	LDA GMM	MEEI, vocal sostenida (53 + 657)	95,9 97,7
Jitter/shimmer Jitter/HNR Shimmer/HNR RPDE no lineal/DFA [17]	QDA	MEEI, vocal sostenida (53 + 632) (53 + 631) (53 + 631) (53 + 654)	81.4 76.4 80.7 91.8
LPC [23]	LDA 3-NN	MEEI, vocal sostenida (42 + 42)	73.0 76.0
Nonlinear HOS + parámetros clásicos Parámetros clásicos [26]	ANN	Privada, vocal sostenida (100 + 68)	98.3 94.4
CHNR [30]	HMM LDA	MEEI, palabra (36 + 607)	65 57
NNE [37]		Privada, vocal sostenida (186 + 64)	78.6
MFCC [20]	LVQ	MEEI, vocal sostenida (53 + 82)	96.0
MFCC + Pitch [21]	HMM	MEEI, vocal sostenida (53 + 657) Habla continua (53 + 657)	98.3 97.8
MFCC + HNR + NNE + GNE + energía [19]	MLP	MEEI, vocal sostenida (53 + 77)	95.1
STFT [31]	SOM	Privada, palabra (18 + 40)	94.8
Wavelets [35]	SVM	Privada, palabras (30 + 30)	98.7
Wavelets [38]	ANN	Privada, vocal sostenida (13 + 51)	84.3
Adaptive TF decomposition [39]	LDA	MEEI, habla continua (51 + 161)	93.4
EMD [32]	3-NN	MEEI, vocal sostenida (53 + 53)	93.4
AD [38]	LDA	MEEI, habla continua (51 + 61)	97.5
Cuantificación TFM [28]	K-means clustering	MEEI, habla continua (51 + 161)	98.6

Tabla 1. Tasas de precisión logradas en diversos trabajos en función del tipo de parametrización, clasificador y base de datos. Adaptada de [40]*. Siglas: AD, ambiguity domain; ANN, artificial neural network; EMD, empirical mode decomposition; GMM, Gaussian mixture model; GNE, Glottal-to-noise excitation; HMM, hidden Markov method; HNR: harmonics-to-noise ratio; LDA, linear discriminant analysis; LPC, linear predictive coding; LVQ, learning vector quantifier; MDVP, multidimensional voice program; MEEI, Massachusetts Eye and Ear Infirmary; MFCC, Mel-frequency cepstral coefficients; MLP, multilayer perceptions; NN, nearest neighbor; NNE, normalized noise energy; SOM, self-organizing map; STFT, short-time Fourier transform; SVM, support vector machine; TFM, time-frequency matrix; RPDE: recurrence period density entropy; DFA: detrended fluctuation analysis; HOS: Higher Order Statistic;*

El mismo estudio [15] también clasifica distintas voces según el tipo de patología, partiendo de diversos grupos de registros de audio que contienen sólo dos patologías distintas. En él se compara el uso de MFCC y EM para realizar tal fin y se obtienen resultados de más del 95% en algunos casos, siendo por norma general más favorable el uso de MFCC. En [25] se evalúa la capacidad de clasificación de enfermedades de la voz proporcionado por el parámetro Glottal to Noise Excitation Ratio (GNE). En general este parámetro resulta adecuado para detectar la existencia o no de una patología si bien no da grandes resultados a la hora de clasificar dicha patología.

En referencia a estudios que se centran en la obtención de un nivel que cuantifique la calidad de la voz, en [41] se introduce un nuevo índice que mide el grado de normalidad en la voz, denominado Pathological Likelihood Index (PLI). Este índice demuestra una alta correlación con la evaluación de la severidad de la patología hecha por un profesional médico (siguiendo la escala GRBAS). Para ello, se segmenta el audio y se parametriza creando vectores MFCC y derivadas relacionadas con variaciones temporales de las características de la señal. Después se crean dos modelos estadísticos distintos (basados en la probabilidad de la existencia de patología o voz normal) y se calcula para cada vector la probabilidad de ser de uno de los dos tipos. A través de estas probabilidades se obtiene el índice PLI. Se observa que la correlación entre este parámetro y otros utilizados en estudios previos es más baja que la de dichos parámetros entre sí. Los resultados indican que el índice PLI es más preciso que otros muy utilizados en la detección de patologías.

Por otro lado, en [42] se realiza la detección del parámetro G de GRBAS usando una base de datos propia en la que sólo aparecen mujeres y que es etiquetada mediante el consenso de tres especialistas. La parametrización se realiza mediante los coeficientes Linear Frequency Spectrum Coefficients (LFSC). El método de clasificación es GMM y se consiguen unos resultados de hasta el 85% de precisión. Estos mismo autores realizan en [43] y [44] un estudio de los distintos márgenes frecuenciales que juegan un papel relevante en la detección del parámetro G de GRBAS concluyendo que la banda de frecuencias entre 0 y 3000 Hz parece ser óptima para la detección de dicho rasgo.

La publicación [45] servirá de punto de partida para este proyecto, ya que también busca realizar una evaluación de GRBAS automática con la misma base de datos que se usará en nuestro estudio. En este caso se utilizan grabaciones (vocal /a/ sostenida y una frase corta) de la base de datos del Hospital Príncipe de Asturias. Cada uno de los registros de audio fue previamente etiquetado por el consenso de tres especialistas para evitar dentro de lo posible la variabilidad inter-evaluador. Este

trabajo utiliza parametrización MFCC de la base de datos junto con la primera y segunda derivada (MFCC + Δ + ΔΔ) y un clasificador tipo Learning Vector Quantization (LQV) obteniendo unos resultados de precisión del 68% para el rasgo G, del 63% para R y valores inferiores para el resto de parámetros.

Debido a que partimos de la misma base de datos y etiquetado de [45], utilizaremos este estudio como línea de base para comparar su precisión con la de un nuevo sistema de parametrización y clasificación diseñado.

4. Objetivos e Hipótesis

En este capítulo se exponen los objetivos del proyecto y la hipótesis sobre la cual se asienta el trabajo realizado.

4.1. Objetivos

El principal objetivo de este proyecto es analizar nuevos mecanismos de parametrización y clasificación de las señales acústicas de la voz que permitan emular una valoración perceptual de su calidad. Estos nuevos mecanismos podrían ser útiles como herramienta de apoyo en el diagnóstico y evaluación de patologías de la voz, dotando al análisis subjetivo realizado por un experto de una cierta base objetiva.

Para ello se analizarán algunos parámetros representativos de la calidad de la voz dentro de los habitualmente utilizados en otros trabajos de investigación, y posteriormente se propondrá un nuevo método de parametrización.

Se modelizarán varios sistemas de parametrización-clasificación que se adecúen a una determinada escala perceptual en función de lo expuesto anteriormente.

Finalmente, se compararán los resultados obtenidos con los distintos sistemas para poder extraer conclusiones y preparar el trabajo futuro de esta investigación.

Este estudio está centrado en el análisis y detección de los dos primeros rasgos de la escala GRBAS (G y R), identificando cada uno de ellos por separado. Se seleccionan tan solo estos dos indicadores como primeros pasos para un estudio posterior más amplio que incluya los cinco rasgos ya que, según [45], parecen tener la mayor consistencia de todo el conjunto.

4.2. Hipótesis

Basándonos en los trabajos [15] y [16] suponemos que el espectro de modulación contiene información suficiente para ayudar en la clasificación de una base de datos de voces según la escala GRBAS. Para ello el EM debe ser post-procesado con el fin de extraer de él un número reducido de características que permitan la clasificación deseada. En principio, unos parámetros como los centroides de dicho espectro podrían ser válidos para la detección automática de los rasgos G y R. Debido a

que el EM tiene información eminentemente frecuencial, también se analizará el uso de otros parámetros como aquellos relacionados con los coeficientes MFCC, la medida del ruido o la complejidad que podrían complementar la información proporcionada por el EM.

Así mismo, ya que los modelos SVM han demostrado ser un buen sistema de clasificación en el reconocimiento de voz, se considera que pueden resultar idóneos en la fase de clasificación. Es por esto por lo que serán utilizados para la detección de los rasgos.

5. Procedimiento y metodología

En este capítulo se detalla el proceso seguido y los métodos utilizados en las distintas fases. Igualmente se describe la base de datos. Las librerías necesarias para la realización del análisis propuesto han sido desarrolladas especialmente para este proyecto excepto en los casos en los que así se indica.

5.1. Introducción

El procedimiento básico que se seguirá en este proyecto comienza con un preprocesado de la base de datos, que incluye una normalización de los registros de audio, eliminación de tramas iniciales y finales, etc. Después se realizará la parametrización, destinada a extraer las características que se utilizarán en la clasificación mediante SVM. Se utilizarán procesos de *data suffling* [46] para realizar distintos entrenamientos. Posteriormente se calculará la correlación cruzada y se obtendrá una matriz de confusión, resultado de las fases de entrenamiento y validación. En la Figura 9 se observa un diagrama de este proceso.

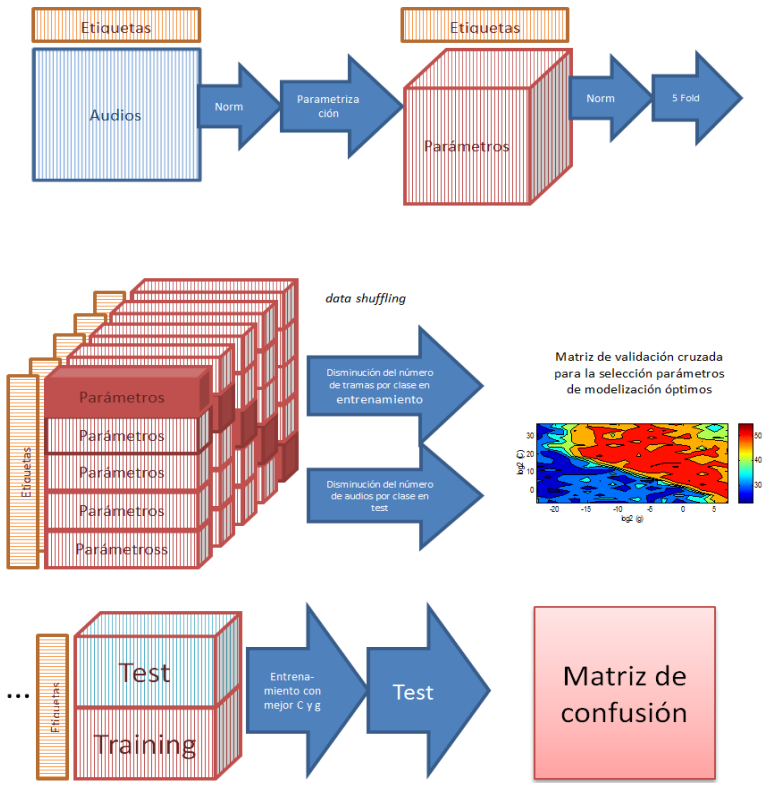


Figura 9. Diagrama de las distintas fases seguidas en este proyecto.

5.2. Segmentación y preprocesado

Todas las grabaciones están realizadas con una frecuencia de muestreo de 50 kHz y 16 bits de cuantificación pero son submuestreadas a 14 kHz tras su paso por un filtro anti-aliasing. La razón de este submuestreo es doble. En primer lugar, se busca una reducción de la cantidad de datos produce más rapidez de parametrización. Por otro lado, y según [42], [43] y [44] la mayor parte de la información relevante a la hora de hacer una evaluación subjetiva por parte de un experto se encuentra debajo de 6 kHz. Otras publicaciones como [47] también demuestran que las voces patológicas tienen una concentración de energía en baja frecuencia.

Cabe destacar que en la publicación [42] se indica que para la clasificación automática del rasgo G se obtiene una máxima precisión al realizar tanto la parametrización como la evaluación del experto en la banda 0-3000 Hz. En nuestro caso la evaluación se hizo en la banda de 0 a 25 kHz ya que la frecuencia de muestreo inicial era de 50 kHz.

En el cálculo de los parámetros se utiliza el espectro desde 0 Hz ya que por debajo de 25 Hz se pueden encontrar características relacionadas con diversas patologías, como enfermedades de carácter neurológico que llevan asociado tremor [3].

Tras el submuestreo, todas las señales son normalizadas en amplitud. No existen silencios en ninguno de los registros de audio ya que son recortados en los periodos transitorios de inicio y final y todos los locutores mantienen la vocal /a/ durante al menos 2 s.

Todos los registros de la base de datos son limitados a la duración del audio más corto y divididos en tramas de distintas longitudes y solapamiento del 50% con el fin de obtener el tamaño de trama óptimo de parametrización. Por norma general la frecuencia fundamental de la voz es pocas veces inferior a 100 Hz lo que implica ciclos de 10 ms, por lo que para tener tramas que contengan al menos tres ciclos estas no deberían tener una duración inferior a 30 ms. También se ha de valorar que muchos de los fenómenos relacionados con algunas patologías, como el tremor, tienen componentes frecuenciales características que comienzan en 4 Hz [18]. De este modo, podría ocurrir que tramas muy cortas no reflejasen claramente la aparición de esas modulaciones de baja frecuencia. Es por esto que se realiza un barrido usando tramas que van desde los 20 a los 500 ms para encontrar la longitud de trama idónea dentro de cada cálculo de parámetros.

5.3. Parametrización

Serán estudiados distintos tipos de parametrización y se observarán los resultados de clasificación para cada uno de ellos. Posteriormente, se realizará una combinación de los parámetros estudiados con los centroides del espectro de modulación (CEM), para observar el incremento en la precisión que pudiera producirse durante la clasificación respecto al caso del uso de un solo tipo de parametrización.

Los cuatro grupos de características utilizadas serán: Centroides de Espectro de Modulación, MFCC, Ruido y Complejidad.

Se seleccionan estos cuatro conjuntos de parámetros ya que, según estudios como [40], es recomendable la combinación de características de distinta naturaleza en la clasificación de voces patológicas. Esto es debido a que ninguna de ellas logra por sí misma caracterizar completamente cada una de las voces. Igualmente, en [26] se indica que los parámetros que deberían ser incluidos para evaluar la calidad de la voz deben contener información acerca de la periodicidad y estabilidad de la voz, la presencia de ruido, la riqueza espectral y los comportamientos no lineales.

Todos los registros presentes en la base de datos serán divididos en tramas de longitud fija con un determinado solapamiento. Sobre cada una de estas tramas se calcularán los parámetros, generándose para cada audio tantos vectores de parámetros como tramas resulten. Siempre se eliminarán la primera y última trama de una parametrización para evitar que ciertos periodos transitorios puedan influir negativamente en la obtención de resultados.

A continuación se desarrollan las cuatro parametrizaciones propuestas.

5.3.1. Centroides de Espectro de Modulación

El espectro de modulación proporciona la información sobre la energía de las frecuencias moduladoras de las portadoras de una señal, siendo una representación bidimensional sobre un eje acústico y otro de modulación frecuencial.

El cálculo del EM se realiza mediante el paso de la señal por un banco de filtros implementado mediante sTFT (short-Time Fourier Transform) de cuyas salidas se detecta la amplitud y envolvente y realiza un análisis frecuencial mediante Discrete Fourier Transform (DFT) [48]. El diagrama de este proceso se muestra en la Figura 10.

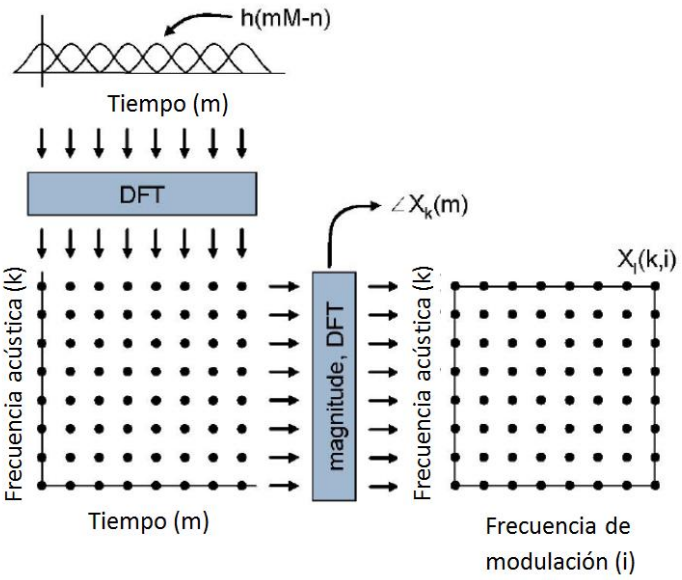


Figura 10. Procedimiento de cálculo del EM [48].

Así, si la señal de entrada es $x(n)$, calculamos la sTFT inicial que determina el eje acústico como:

$$X_k(m)= \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_K^{kn} \qquad \text{[ec. 1]}$$

para $k = 0, \dots, K-1$

Y la detección de envolvente y análisis frecuencial posterior que determina el eje de frecuencia de modulación como:

$$X_l(k,i)= \sum_{m=-\infty}^{\infty} g(IL-m)|X_k(m)|W_l^{im} \qquad \text{[ec. 2]}$$

para $i=0,\dots,I-1$

siendo $W_K = e^{-j(\frac{2\pi}{K})}$ [ec. 3]

En donde, $h(n)$ es la ventana de análisis del eje acústico y $g(n)$ la del eje de modulación, k representa el número de banda acústica e i el de modulación. K e I

representan el número total de líneas que existirán en el eje acústico y de modulación respectivamente.

Es decir, en primer lugar se realiza una sTFT de la señal temporal dando lugar a un espectograma. Posteriormente se realiza una segunda sTFT sobre el eje temporal de dicho espectrograma dando lugar al espectro de modulación.

Este EM nos permite observar varias características de la voz de forma simultánea, como su naturaleza armónica, a la vez que pueden verse modulaciones de la frecuencia fundamental y de alguno de estos armónicos.

Este tipo de parametrización se utiliza en [15] y [16] para la clasificación de los registros de audio de la base de datos PdA como normales/patológicos y para la clasificación automática de algunas enfermedades en [49], [50], [43] y [44], así como en [15] para la obtención de parámetros objetivos que permitan cuantificar la calidad de la voz. En [53] se indica que el EM de pacientes con patologías no tiene picos tan armónicos, y en general es más aplanado y disperso que el de individuos sin problemas de fonación. En ese estudio, la parametrización se agrupan los registros de audio en función de la patología y se usan criterios de Máxima Relevancia [54] basados en las clases existentes para la extracción de características. Es decir, partiendo de los EM de las distintas clases de las que disponen, un algoritmo extrae los parámetros más representativos de cada una de esas clases para ser utilizados en el proceso de clasificación.

Se ha de tener en cuenta que pretendemos extraer características relevantes a partir del EM que tengan una cierta correlación con la evaluación perceptual realizada por un especialista. Estudios como [55] y [56] hablan de la relación no lineal que hay entre los estímulos físicos de la voz y las variaciones en la señal acústica y una evaluación o percepción subjetiva de esta. El EM nos permite conocer la distribución energética del espectro de modulación frente al acústico. Por ejemplo, podremos saber a qué frecuencias se modula el pitch de una voz o uno de sus armónicos, lo que podría ser relevante desde el punto de vista de la parametrización de distintas patologías; si bien es cierto que según [57] los especialistas médicos perciben de una forma muy distinta las modulaciones en función de si estas son sinusoidales o no, lo que conlleva un etiquetando diferente para cada uno de los registros de audio. Puede ocurrir que dos registros que tienen un EM similar contengan señales de modulación con el mismo ancho espectral y frecuencia fundamental pero distinta forma de onda, lo que producirá una percepción diferente en el experto evaluador. Es por esto que podríamos pensar que es posible que el EM no contenga suficiente información como para proporcionar parámetros únicos en la clasificación y tal vez deba usarse como un

complemento a otros parámetros. Del mismo modo, en [58] se observa que existen muchos parámetros cuantificables que podrían no verse reflejados en un EM y que sí tienen un carácter perceptivo de cara a un experto.

En la Figura 11 podemos observar el espectro de modulación de dos señales con la misma portadora y distintas modulaciones

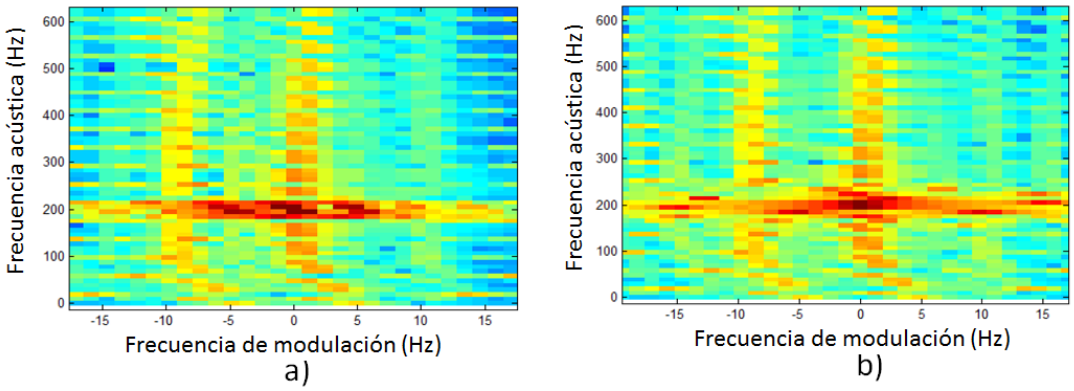


Figura 11. a) Espectro de modulación de una sinusoide de 200 Hz con modulación de amplitud a 5 Hz. b) Espectro de modulación de una sinusoide de 200 Hz con una modulación en frecuencia a 15 Hz y profundidad de 8 Hz.

Para el cálculo del espectro de modulación se utiliza la librería Modulation Toolbox ver. 2.1 [59]. Este toolbox permite realizar el EM de una señal teniendo como variables de entrada los siguientes parámetros:

- Opciones de demodulación: COG (Center Of Gravity), Harmonic, Hilbert, Harm-COG
- Número de bandas acústicas y ancho de banda de cada una de ellas
- Número de líneas del eje de modulación
- Tipo de enventanado: Rectangular, Hanning, Hamming…

El resultado es un espectro con tantas bandas acústicas como las indicadas a la entrada, cada una con su propio ancho de banda y tantas líneas de modulación como se especifiquen. Cada una de las líneas equivaldrá a un ancho de banda fijo que será igual a la frecuencia de muestreo dividida entre el número de líneas.

Por un lado, se parametrizará el espectro de modulación con un determinado número de bandas acústicas que oscilará normalmente entre 30 y 90 . La distancia

entre las frecuencias inferior (f_1) y superior (f_2) de cada banda se calculará del siguiente modo:

$$f_2 = \sqrt[a]{2} \cdot f_1 \qquad \text{[ec. 4]}$$

$$\text{Con } a = \frac{\log_2\left(\frac{f_{máx}}{50}\right)}{N} \qquad \text{[ec. 5]}$$

Siendo $f_{máx}$ la frecuencia máxima deseada y N el número de bandas.

Algo similar se hará con el eje de frecuencia de modulación en donde se desea tener bandas con ancho de banda percentual en fracción de octava que reduzca el número de bandas a un valor normalmente inferior a 40.

Tras el cálculo del EM se debe extaer una determinada cantidad de información representativa de este para ser utilizado en la etapa de clasificación. Se decide utilizar el cálculo de centroides como método de extracción de características del EM ya que además de proporcionar un número reducido de datos, lo cual es positivo para la fase de clasificación, proporcionan una información muy visual sobre las frecuencias en las que se distribuye la energía en el espectro. Trabajos como [60] utilizan los centroides del espectro de audio en reconocimiento de voz. Si se confirma que esta información es útil para la evaluación automática de la calidad de la voz, los centroides podrán usarse con una doble función, ya que muestran de una forma sencilla y visual información relevante para el especialista médico. Por lo tanto estos podrían apoyar su diagnóstico en la observación de los centroides, más sencillos de interpretar que un espectro de modulación o un espectrograma.

El cálculo del centroide es simple. Sólo se debe multiplicar cada frecuencia por el valor de la energía en esa línea frecuencial, sumar todos los valores y dividir entre la suma energética total:

$$\text{Centroide}(i) = \frac{\sum_{k=0}^{K-1} \text{Frec}(k)E(k)}{\sum_{k=0}^{K-1} E(k)} \qquad \text{[ec. 6]}$$

Siendo k e i las línea frecuenciales del eje acústico y de modulación respectivamente, K el número total de líneas, $\text{Frec}(k)$ el valor frecuencial (Hz) de la línea k y $E(k)$ su energía.

Con el fin de realizar una normalización respecto a la frecuencia fundamental, todos los centroides resultantes son divididos por el valor global (f_0) de dicha frecuencia a lo largo de todas las tramas.

Por lo tanto, tendremos:

$$\text{Centroide(i)} = \frac{\sum_{k=0}^{K-1} \text{Frec}(k)E(k)}{f_0 \sum_{k=0}^{K-1} E(k)} \qquad \text{[ec. 7]}$$

En la Figura 12 se observan los espectros de modulación y centroides de distintos tipos de voces, en los que se pueden ver las diferencias entre la voz normal y las patológicas.

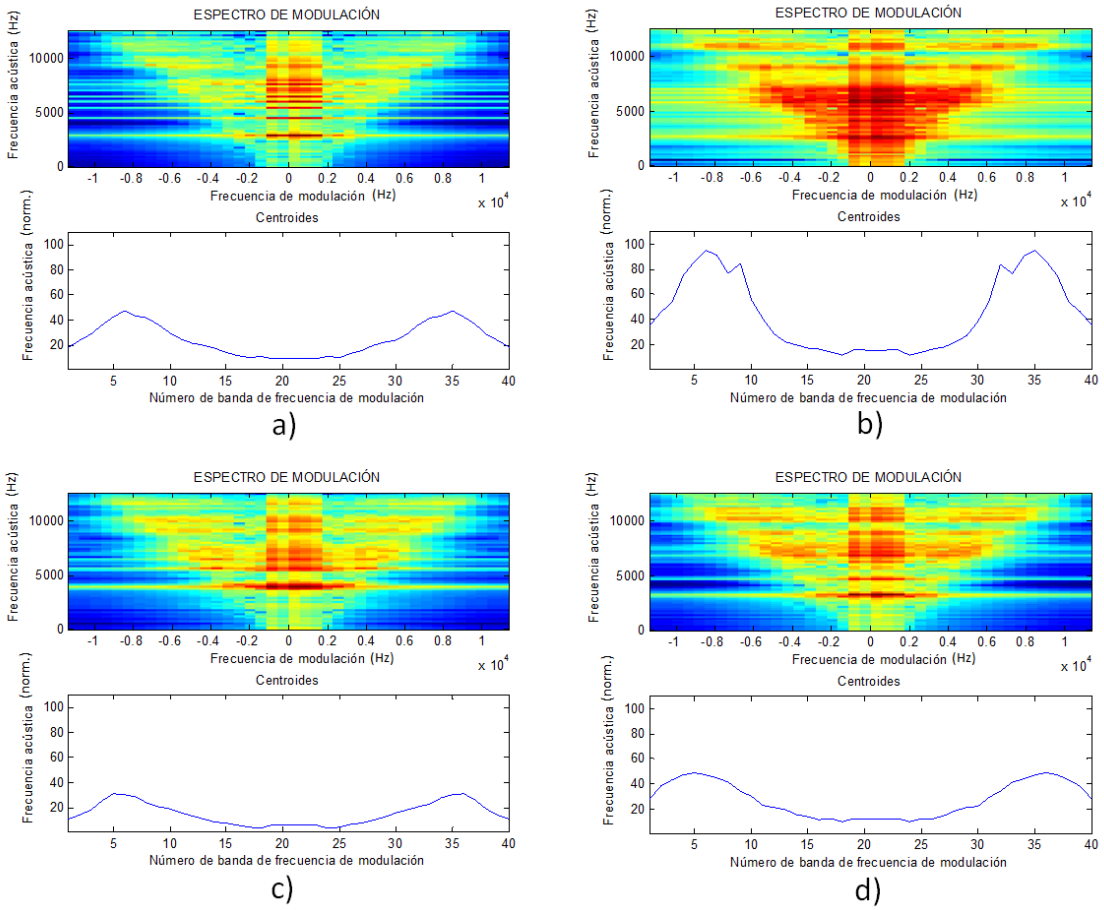


Figura 12. EM y centroides normalizados de distintos tipos de voces. a) Voz normal; b) Laringitis crónica hiperplástica; c) Lesión de neurona motora superior; d) Quiste epidermoide

En función de todo lo expuesto anteriormente, las distintas variables que se utilizan para obtener los centroides serán:

- Número de bandas acústicas
- Número de bandas de frecuencia de modulación (número de centroides)

Se realizará un barrido de estos parámetros combinados con los citados para el cálculo del EM, para distintas longitudes de trama, hasta encontrar la parametrización que arroje unos resultados de clasificación óptimos.

5.3.2. Mel Frequency Cepstral Coefficients (MFCC)

Otro conjunto de parámetros a tener en cuenta y muy utilizado en la detección de patologías de la voz son los coeficientes Mel-Frequency Cepstral.

Estos coeficientes suelen ser utilizados en multitud de aplicaciones relacionadas con el habla. Representan la energía distribuida en bandas sobre una escala frecuencial perceptual relacionada con el sistema auditivo humano denominada escala Mel [22]. De este modo permiten identificar ciertos aspectos relacionados con la percepción del habla, siendo altamente útiles en aplicaciones de reconocimiento de locutor o detección de patologías.

Para calcular los parámetros MFCC (c_m) se divide la energía de la señal de entrada $x(n)$ en bandas mediante un enventanado triangular en el dominio de la frecuencia:

$$S_p = \sum_{q=0}^{\frac{p}{2}-1} W_p(q) \cdot X(q) \qquad \text{[ec. 8]}$$

$$\text{Con } q = 1, \dots, Q$$

Siendo S_p la energía de cada banda, Wp la función de enventanado y Q el número de bandas mel. Posteriormente, se obtienen los coeficientes realizando la transformada discreta del coseno al logaritmo de S_p.

$$c_m = \sum_{p=1}^{M} \log(S_p) \cdot \cos\left[n \cdot (p-0.5) \cdot \frac{\pi}{M}\right] \qquad \text{[ec. 9]}$$

El proceso puede esquematizarse según se muestra en la Figura 13.

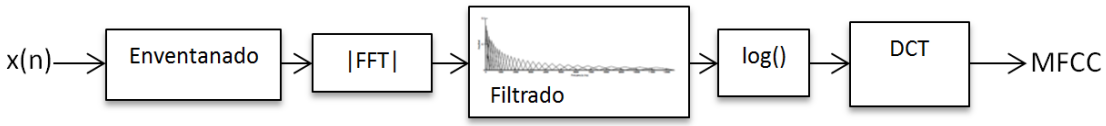


Figura 13. Proceso de obtención de coeficientes MFCC

Los coeficientes MFCC ya han sido utilizados en la clasificación de los rasgos GRBAS en [45]. En este proyecto intentaremos, por un lado, reproducir los resultados obtenidos para los rasgos G y R, y por otro, comprobaremos el efecto en la clasificación que supone añadir los centroides del EM a estas características. En dicho proyecto se usan los parámetros MFCC combinados con su primera y segunda derivada (Δ y ΔΔ). Si $c_m(t)$ son los coeficientes en el instante t, tendremos:

$$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx \mu \sum_{q=-R}^{R} q \cdot c_m(t+q) \qquad \text{[ec. 10]}$$

Siendo μ una constante de normalización y $(2R+1)$ el número de tramas sobre las que se realiza el cálculo.

Para el cálculo de MFCC se utilizará la librería desarrollada en el laboratorio BYO del departamento ICS de la EUITT para Matlab® destinada al cálculo de parámetros de calidad de la voz.

5.3.3. Parámetros de complejidad

Los parámetros de complejidad aportan información cuantitativa sobre la no linealidad en el funcionamiento de las cuerdas vocales y del aparato fonador en general, muchas de las cuales pueden estar provocadas por diversas patologías. En muchos casos resultan ser un buen indicador de la presencia de una determinada disfunción, lo cual ha sido objeto de estudio en [21] y [61]. Este tipo de medidas han demostrado ser más eficaces en la detección y clasificación de voces patológicas con importantes modulaciones y subarmónicos que las de jitter y shimmer [62].

Para la caracterización de los registros de audio de la base de datos se utilizarán los siguientes parámetros:

- Largest Lyapunov Exponent (LLE). En líneas generales indica el nivel de estabilidad del sistema [63].

- Correlation Dimension (CD). Proporciona una Buena medida de la complejidad de la dinámica. Mide el número de grados de libertad activos [64].

- Aproximate Entropy (AE). Partiendo de que la entropía cuantifica la incertidumbre de una variable aleatoria, la entropía aproximada mide el promedio de la información condicional generada por puntos divergentes de la trayectoria [65], [66].

- Sampled Entropy (SE). Es muy similar a la Entropía Aproximada pero proporciona un valor menos dependiente de la longitud de la señal [67].

- Modified sampled entropy (MSE). Es muy similar a la Sampled Entropy pero utiliza una función sigmoide no lineal para su cálculo, es más robusta al ruido, presenta mayor consistencia relativa y es aún menos dependiente de la longitud de la señal [68].

- Gaussian Kernel Aproximate Entropy (GKAE). Es muy similar a la entropía aproximada pero se calcula con un algoritmo que la hace más consistente, estable y precisa estadísticamente [69].

- Fuzzy Entropy (FE). Se trata de una entropía no probabilística que permite resolver alguno de los problemas de la teoría de Bayes [70].

- Detrended Fluctuation Analysis (DFA). Proporciona información sobre los cambios en el ruido de respiración de la voz. Además, su cálculo no depende del pitch de esta por lo que es útil para voces patológicas ya que suele ser difícil detectar la frecuencia fundamental en este tipo de señales [17]

- Recurrence Period Density Entropy (RPDE). Mide y cuantifica la existencia de alguna ambigüedad en la frecuencia fundamental, lo cual suele ser un indicativo de la existencia de una patología [71].

Para el cálculo de los parámetros de complejidad se utilizará la librería desarrollada en el laboratorio BYO del departamento ICS de la EUITT para Matlab ® destinada al cálculo de parámetros de calidad de la voz.

5.3.4. Parámetros de medida de ruido

Debido a que en presencia de muchas patologías aparecen ciertas componentes no armónicas en la señal de la voz, la medida de ruido es utilizada frecuentemente

como fuente de datos objetivos para caracterizar la calidad de la voz. Varios trabajos utilizan estas medidas para detección de patologías como [72], [73] y [74].

Por lo tanto, como cuarto método de parametrización se utilizarán medidas de ruido de la voz. Para cada trama se calcularán:

- Harmonics to Noise Ratio (HNR): Relación entre la energía de los armónicos y la del ruido para cuantificar la sensación perceptual de disfonía. [75]
- Normaized Noise Energy (NNE): Proporciona información referente al nivel del ruido en la señal analizada respecto al nivel total de dicha señal. Para ello se necesita calcular el pitch de la voz y detectar los picos de nivel [37].
- Glottal to Noise Excitation Ratio (GNE): Se define como la relación entre la cantidad de señal debida a las cuerdas vocales y el ruido producido por el aire que escapa entre estas y no influye directamente en la fonación [76].
- Voice Turbulence Index (VTI): Está correlacionado con las turbulencias causadas por los fallos en la aducción de las cuerdas vocales [77].
- Soft Phonation Index (SPI): Indica la falta de componentes de alta frecuencia en los armónicos de la voz, lo que podría ser indicativo de un mal funcionamiento de las cuerdas vocales [77].
- Cepstrum based Harmonics to Noise Ratio (CHNR). Realiza el cálculo de HNR en el dominio cepstrum siendo más preciso que HNR en presencia de jitter [78].
- Noise to harmonics ratio (NHR): Relación ruido-armónicos en unidades lineales [79].

Para el cálculo de los parámetros de ruido se utilizará la librería desarrollada en el laboratorio BYO del departamento ICS de la EUITT para Matlab® destinada al cálculo de parámetros de calidad de la voz.

5.4. Post-procesado

Tras la obtención de todos los parámetros estos deberán ser procesados para poder realizar una correcta clasificación. El post-procesado que se realiza sobre ellos es muy sencillo y consiste en normalización y fusión de características, si procede.

- Normalización: cada conjunto de características se normaliza para que el valor máximo sea 1 y el mínimo sea -1 si los parámetros tienen valores positivos y negativos o 0 si sólo existen valores positivos. Este tipo de normalización es un

requerimiento de la librería utilizada para el entrenamiento de los sistemas SVM.

- Fusión de características: inicialmente se analizará el comportamiento de cada una de las familias de parámetros propuestas para la obtención de un sistema de clasificación que simule una evaluación perceptual, y posteriormente se repetirá el proceso utilizando una fusión de características ya normalizadas. De este modo se podrá comprobar si la información de unos parámetros resulta complementaria a la de otros, y por lo tanto se produce una mejora en el sistema. En este proyecto sólo se combinarán características que tengan la misma longitud de trama.

5.5. Clasificación

En esta etapa se utilizará un clasificador del tipo Support Vector Machines [33], [34]. Básicamente se puede afirmar que los algoritmos de aprendizaje supervisado SVM crean un espacio de múltiples dimensiones en el que se sitúan los vectores de características o puntos de entrada, de tal forma que se intenta que en dicho espacio las clases estén lo más separadas posible a través de hiperplanos. Así, cuanto mayor sea el espacio entre dos clases (o mayor sea la distancia de un hiperplano a la case más cercana), mejor será la generalización. De este modo, cuando queramos clasificar un vector de clase desconocida, lo introduciremos en el modelo de clasificación SVM ya calculado que lo situará en una determinada región. Posteriormente se etiquetará como perteneciente a la clase representada por dicha región. Puede verse un ejemplo gráfico de espacio de dos clases e hiperplano en la Figura 14.

Los algoritmos SVM utilizan un *kernel* o función núcleo que permite transformar el espacio inicial de características en uno multidimensional más complejo que proporcione una separación de clases adecuada. Podemos encontrar kernels lineales, de tipo polinomial, sigmoideo o de base radial gausiana entre muchos otros. En el caso de este proyecto se realizarán entrenamientos únicamente con funciones del tipo base radial gausiana. Este kernel consiste en una función exponencial en cuyo exponente aparece la variable numérica g. Por lo tanto existirán dos posibles variables a la hora de entrenar un clasificador SVM: C y g.

- g: es la variable de la función exponencial del kernel.

- C: permite controlar la frontera entre una clase y otra proporcionando un “soft margin” que consiente ciertos errores de clasificación en la fase de entrenamiento. De esta forma se evita el sobreajuste. El sobreajuste se produce

cuando el modelo generado se ajusta tanto a los vectores que se han utilizado para su entrenamiento que no representa bien el caso general. En estas ocasiones aumenta la posibilidad de que un nuevo vector de clase desconocida sea clasificado erróneamente si cae cerca de una frontera o hiperplano. La Figura 15 ilustra gráficamente un ejemplo de sobreajuste y otro de generalización.

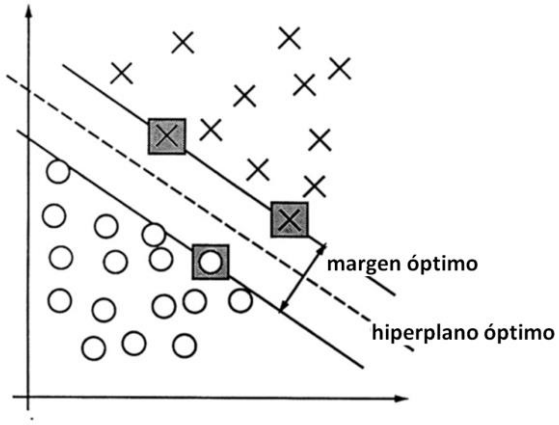


Figura 14. Ejemplo de espacio conteniendo dos clases. Los vectores de soporte marcados en gris son los que determinan el margen entre dos clases. Adaptado de [33].

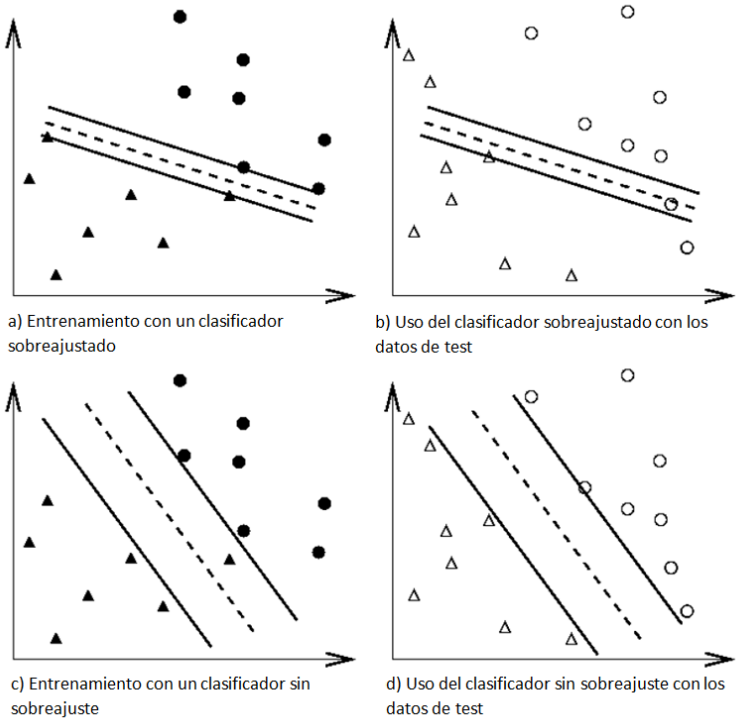


Figura 15. Ejemplo de clasificador SVM de dos dimensiones con sobreajuste (a y b) y sin sobreajuste (c y d). Adaptado de [80].

Tras la parametrización se entrenará el clasificador SVM para un rango de valores de C $\{2^{-7}, 2^{37}\}$ y g $\{2^{-23}, 2^{7}\}$ con intervalos de 2^2. Debido al desbalance existente entre las distintas clases para los rasgos G y R no se seleccionarán todas las tramas disponibles para el entrenamiento. En el análisis comparativo con [45] se usarán 1000 tramas por clase pero en el resto de entrenamientos se limitará a 600 el máximo número de tramas por cada clase. Estas se extraerán aleatoriamente de todos los posibles pacientes para así aumentar la variabilidad inter-locutor. Para algunas longitudes de trama y la cuarta clase (nivel 3) en algunos casos no se llegará a tener 600 tramas como puede deducirse de los histogramas de cada uno de los rasgos, representado en la Figura 16.

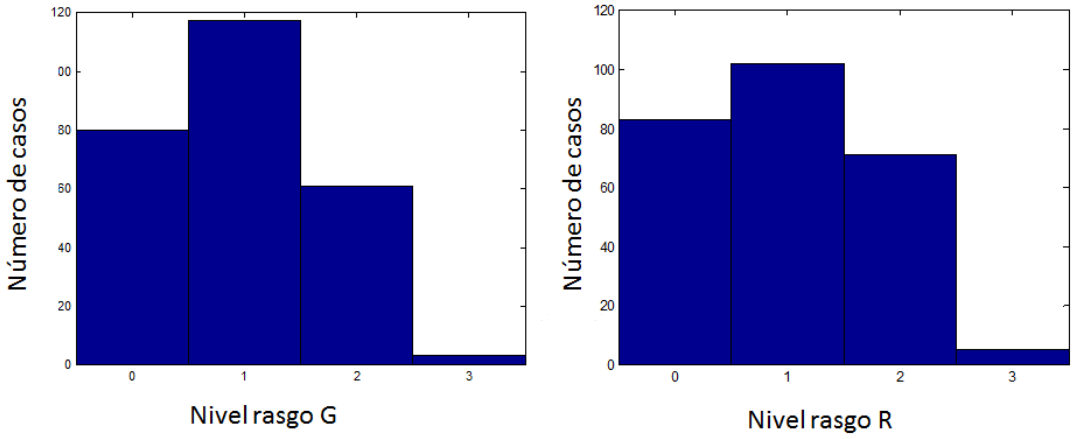


Figura 16. Histograma de los cuatro niveles en cada uno de los rangos estudiados.

Para efectuar el entrenamiento y la validación se utilizan algunas de las funciones de la librería Libsvm [81].

5.5.1. Validación cruzada con K-folds

Para realizar la validación cruzada se utilizará una técnica denominada *data suffling* mediante la cual se divide la base de datos completa en cinco capas, de las cuales cuatro son utilizadas para entrenar un modelo con unos determinados valores de C y g. La capa restante se destina a la validación del modelo calculado. Este proceso se repite otras cuatro veces más dejando siempre una capa distinta para la validación y almacenando el resultado de la precisión obtenida en cada caso. Al final de todo el proceso se observa la matriz con los valores medios de precisión obtenidos para cada C y g, seleccionándose los valores que proporcionen una precisión máxima, obteniéndose así la validación cruzada [36], [82]. Posteriormente, se realiza un

segundo barrido más fino en el cual se repite el proceso usando valores de C y g focalizados alrededor de los citados anteriormente y con intervalos de prueba más pequeños. De esta forma se obtendrán unos valores de C y g óptimos con los que se calculará un nuevo modelo partiendo del 70% de los registros de audio, seleccionados aleatoriamente (conservando el balance de clases inicial de la base de datos). Finalmente, se realizará una validación con el 30% de audios restantes obteniéndose así la matriz de confusión.

5.6. Material de partida. Base de Datos

Partimos de la base de datos realizada en el Hospital Príncipe de Asturias de Alcalá de Henares (PdA) de la que utilizaremos la grabación de la vocal /a/ sostenida de 413 voces normales y 200 patológicas si bien para algunas pruebas se reducirá el número de voces normales a 76 y patológicas a 185 con el fin de balancear las clases existentes (de otra forma el número de repeticiones de la clase representativa del valor 0 sería muy elevado). De este modo también se agilizan las comparaciones entre las distintas parametrizaciones que se llevan a cabo. Todos los archivos de audio son mono y tienen una duración aproximada de 2s siendo registrados en las mismas condiciones. En concreto, las grabaciones se realizaron con el sistema de adquisición CLS 4300B de Kay Elemetrics, con frecuencia de muestreo de 50 kHz y 16 bits de cuantificación. Se utilizó un micrófono de condensador situado a 30 cm de la boca de los locutores y con un ángulo de 50º. Todas las muestras se toman en una sala acústicamente aislada [83].

Las voces normales utilizadas cumplen con los siguientes criterios de inclusión:

- El locutor no percibe subjetivamente ningún problema en su laringe.
- Voz adecuada para su edad, género y grupo cultural con un pitch, volumen, timbre y dicción adecuados.
- No fumador.
- Nunca ha sido intervenido clínicamente por patología en la laringe.
- No ha requerido entubación endotraqueal durante el último año.
- Un nivel GRBAS inferior a 3.

Todos los locutores clasificados como normales fueron explorados con un laringoestroboscopio para descartar la existencia de problemas en la laringe.

En cuanto a las voces patológicas, fueron registradas en las visitas de los pacientes al Hospital Príncipe de Asturias y cubren un amplio rango de patologías.

A modo de resumen, el **corpus completo** tiene las siguientes estadísticas, indicadas en las Tablas 2, 3 y 4:

	Voces normales			**Voces patológicas**		
Edades	Hombres (142 audios)	Mujeres (271 audios)	Total (413 audios)	Hombres (74 audios)	Mujeres (126 audios)	Total (200 audios)
Mínima	18.0	13.0	13.0	11.0	9.0	9.0
Máxima	74.0	66.0	74.0	76.0	72.0	76.0
Media	31.5	33.3	32.5	48.1	36.6	40.9
Desviación	13.7	13.1	13.4	13.9	13.0	14.4

Tabla 2. Estadística de edades del corpus completo

Patología presente	**Número de casos**
Sulcus	1
Sulcus en estría	22
Quiste epidermoide	20
Adquiridas traumáticas iatrógenas sobre las cuerdas vocales	2
Laringitis crónica hiperplásica	19
Laringitis crónica hiperplásica con leucoplasia	11
Parálisis periféricas	1
Parálisis periféricas: Recurrente derecho	9
Parálisis periféricas: Recurrente izquierdo	8
Lesión de neurona motora superior	14
Alteraciones extrapiramidales	1
Nódulo bilateral	29
Pólipo pediculado	28
Edema de Reinke bilateral	29
Falta de cierre	6
Total	200

Tabla 3. Frecuencia de aparición de patologías en el corpus completo

Frecuencia de aparición		
Nivel	Rasgo G	Rasgo R
0	396	381
1	141	143
2	71	82
3	5	7
Total	613	613

Tabla 4. Frecuencia de aparición de cada uno de los niveles para ambos rasgos en el corpus completo

Y para el corpus reducido, las estadísticas se indican en las Tablas 5, 6 y 7:

	Voces normales			**Voces patológicas**		
Edades	Hombres (34 audios)	Mujeres (42 audios)	Total (76 audios)	Hombres (71 audios)	Mujeres (114 audios)	Total (185 audios)
Mínima	19.0	13.0	13.0	11.0	9.0	9.0
Máxima	74.0	66.0	74.0	76.0	72.0	76.0
Media	30.7	31.0	30.9	45.3	34.1	38.1
Desviación	14.1	14.1	14.1	13.7	12.6	14.1

Tabla 5. Estadística de edades del corpus reducido

Patología presente	**Número de casos**
Sulcus	1
Sulcus en estría	21
Quiste epidermoide	19
Adquiridas traumáticas iatrógenas sobre las cuerdas vocales	2
Laringitis crónica hiperplásica	17
Laringitis crónica hiperplásica con leucoplasia	10
Parálisis periféricas	1
Parálisis periféricas: Recurrente derecho	9
Parálisis periféricas: Recurrente izquierdo	6
Lesión de neurona motora superior	8
Alteraciones extrapiramidales	1
Nódulo bilateral	29
Pólipo pediculado	28
Edema de Reinke bilateral	28
Falta de cierre	5
Total	185

Tabla 6. Frecuencia de aparición de patologías en el corpus reducido

Frecuencia de aparición		
Clase	Rasgo G	Rasgo R
0	80	83
1	117	102
2	61	71
3	3	5
Total	261	261

Tabla 7. Frecuencia de aparición de cada uno de los niveles para ambos rasgos en el corpus reducido

El corpus completo se utilizará para clasificar mediante SVM toda la base de datos y comparar los resultados con los obtenidos en [45]. Posteriormente, se usará la base de datos reducida para comparar los resultados obtenidos con MFCC sobre esta base de datos respecto a la realizada en [45]. Por último se realizarán todas las parametrizaciones propuestas y se compararán los resultados entre sí.

Cada uno de los registros de audio de esta base de datos ha sido etiquetado según la escala GRBAS por el consenso de tres expertos en una única sesión.

6. Pruebas y resultados

A continuación se muestran los resultados de eficiencia media tras la validación cruzada y las matrices de confusión obtenidos con cada una de las parametrizaciones por separado y los de la fusión de CEM con el resto.

6.1. Uso de Centroides de EM

Como ya se había indicado, se pre-procesa la base de datos de 261 registros de audio y se calculan los centroides del EM para distintos valores de las variables de parametrización, obteniéndose los mejores resultados para una longitud de trama de 100 ms, 26 centroides, 1024 líneas en el eje de frecuencia de modulación y 70 bandas acústicas. La Tabla 8 muestra la matriz de confusión para este caso.

		G				R			
		Niveles predichos							
		0	1	2	3	0	1	2	3
Niveles de entrada	0	12	10	0	2	16	7	1	0
	1	3	27	5	0	7	20	3	0
	2	0	11	6	1	2	16	3	0
	3	0	1	0	0	1	0	0	0

Tabla 8. Matriz de confusión para los rasgos G y R con parametrización tipo CEM (100 ms, 26 centroides, 1024 líneas y 70 bandas acústicas).

Para este caso, los resultados de validación cruzada arrojan una eficiencia media del **59%** (varianza: 7%) para el rasgo **G** y del **48 %** (varianza: 5%) para **R**.

Posteriormente se fusionarán las parametrizaciónes de centroides basadas en EM con otros parámetros para comprobar si la combinación de parámetros supone alguna ventaja.

6.2. Uso de coeficientes MFCC

Como se indicó anteriormente, se parte del trabajo [45] donde se realiza un estudio similar al de este proyecto utilizando los coeficientes MFCC+Δ+ΔΔ con 45 parámetros (15 coeficientes MFCC, 15 coeficientes de la primera derivada y 15 de la segunda) sobre tramas de 40 ms para detectar los rasgos GRBAS. En este proyecto se utiliza la misma base de datos con dos modificaciones:

- Se excluyen el 5,4% de los registros de audio por presentar problemas relacionados con errores en la grabación o por ser sospechosos de tener un etiquetado erróneo.

- Se utilizan los registros de audio a una frecuencia de muestreo de 14 kHz ya que no se dispone de la base de datos muestreada a 50 kHz.

No se espera que ninguna de las dos modificaciones influya de forma significativa ya que el número de registros eliminado es muy bajo y la señal de la voz tiene la mayor parte de la información por debajo de 7 kHz.

Se intentan reproducir los resultados obtenidos en dicho trabajo pero con el uso de un clasificador SVM en lugar de una red LVQ.

Debido a que los resultados proporcionados por [45] no tienen validación cruzada sino que se han obtenido mediante el entrenamiento de un modelo con el 70% de la base de datos y un 30% para validación se realizan las mismas pruebas cogiendo cinco agrupaciones distintas de los mismos datos con la distribución citada anteriormente. Como puede verse en la Tabla 9, en al menos uno de ellos se logra reproducir un resultado similar.

	Eficiencia			
	Valor máximo		Validación Cruzada	
	G	R	G	R
Clasificación LVQ	68,00%	63,00%	-	-
Clasificación SVM	65,00%	67,00%	57,00%	52,00%

Tabla 9. Comparativa entre valores máximos de eficiencia de clasificación y validación cruzada entre dos clasificadores distintos, LVQ (realizada en [45]*) y SVM (realizada en este proyecto).*

De este modo se puede considerar que ambas clasificaciones arrojan valores similares.

De ahora en adelante todos los resultados y comparativas entre parametrizaciones serán realizados con el corpus reducido, que incluye 261 registros de audio.

En la Tabla 10 se observan los resultados de la parametrización con MFCC+Δ+ΔΔ para una trama de 100 ms y un solapamiento entre tramas del 50%.

		G				R			
		Etiquetas predichas							
		0	1	2	3	0	1	2	3
Niveles de entrada	0	12	12	0	0	6	17	1	0
	1	8	24	2	1	6	23	1	0
	2	3	9	6	0	0	16	3	2
	3	1	0	0	0	0	0	1	0

Tabla 10. Matriz de confusión para los rasgos G y R con parametrización tipo MFCC+Δ+ΔΔ (100 ms).

Los resultados de validación cruzada arrojan una eficiencia media del **51%** (varianza: 7%) para el rasgo **G** y del **48 %** (varianza: 8%) para **R**. Los resultados obtenidos para 40 ms tenían una precisión ligeramente inferior.

Por otro lado, al combinar características de MFCC+Δ+ΔΔ (sobre 15 coeficientes) y centroides de EM (26 centroides), 70 bandas acústicas para tramas de una longitud de 100 ms y solapamiento del 50% se obtienen los resultados indicados en la Tabla 11.

		G				R			
		Niveles predichos							
		0	1	2	3	0	1	2	3
Niveles de entrada	0	14	8	2	0	15	6	3	0
	1	10	23	2	0	12	12	6	0
	2	2	9	7	0	3	8	9	1
	3	0	1	0	0	0	1	0	0

Tabla 11. Matriz de confusión para los rasgos G y R con parametrización MFCC +Δ +ΔΔ y CEM de 26 centroides (100 ms).

Para este caso, los resultados de validación cruzada arrojan una eficiencia media del **55%** (varianza: 6 %) para el rasgo **G** y del **49 %** (varianza: 6 %) para **R**. Si bien para tramas de 80 ms la eficiencia de clasificación es del **56%** para el rasgo **G** y del **51 %** para **R.**

6.3. Uso de parámetros de complejidad

Se parametriza la base de datos para obtener valores de complejidad de cada trama en función de los distintos valores de las variables de entrada, obteniéndose los mejores resultados para una longitud de trama de 100 ms según se muestra en la Tabla 12.

		G				R			
		Niveles predichos							
		0	1	2	3	0	1	2	3
Niveles de entrada	0	21	3	0	0	13	9	2	0
	1	8	20	7	0	8	18	4	0
	2	1	8	9	0	1	11	9	0
	3	0	0	0	1	0	0	1	0

Tabla 12. Matriz de confusión para los rasgos G y R con parametrización de valores de complejidad (100 ms).

Para este caso, los resultados de validación cruzada arrojan una eficiencia media del **60 %** (varianza: 6%) para el rasgo **G** y del **53 %** (varianza: 6%) para **R**.

Por otro lado, al combinar características de valores de complejidad y centroides de EM (26 centroides) para tramas de una longitud de 100 ms y solapamiento del 50% se obtienen los resultados indicados en la Tabla 13.

		G				R			
		Niveles predichos							
		0	1	2	3	0	1	2	3
Niveles de entrada	0	14	10	0	0	13	10	1	0
	1	2	29	4	0	8	18	4	0
	2	0	8	9	1	2	10	9	0
	3	0	0	1	0	0	0	1	0

Tabla 13. Matriz de confusión para los rasgos G y R con parametrización de complejidad y CEM de 26 centroides (100 ms).

Con la combinación de parámetros, los resultados de validación cruzada arrojan una eficiencia media del **59 %** (varianza: 7 %) para el rasgo **G** y del **53 %** (varianza: 5%) para **R**. Entrenamientos realizados con otras variables de parametrización no proporcionan mejores resultados.

6.4. Uso de parámetros de medida de ruido

Se parametriza la base de datos para obtener valores de ruido de cada trama en función de los distintos valores de las variables de entrada, obteniéndose los mejores resultados para una longitud de trama de 100 ms según se muestra en la Tabla 14.

		G				R			
		Niveles predichos							
		0	1	2	3	0	1	2	3
Niveles de entrada	0	8	16	0	0	6	17	1	0
	1	4	31	0	0	6	23	1	0
	2	0	12	6	0	0	16	3	2
	3	0	0	1	0	0	0	1	0

Tabla 14. Matriz de confusión para los rasgos G y R con parametrización de índices de Ruido HNR, NNE, GNE, VTI, SPI, CHNR, NHR (100 ms).

Para este caso, los resultados de validación cruzada arrojan una eficiencia media del **55 %** (varianza: 10 %) para el rasgo **G** y del **49 %** (varianza: 7 %) para

R. Entrenamientos realizados con otras variables de parametrización no proporcionan mejores resultados.

Por otro lado, al combinar características de ruido y centroides de EM (26 centroides) para tramas de una longitud de 100 ms y solapamiento del 50% se obtienen los resultados indicados en la Tabla 15.

		G				R			
		Niveles predichos							
		0	1	2	3	0	1	2	3
Niveles de entrada	0	10	14	0	0	19	4	1	0
	1	4	31	0	0	7	16	7	0
	2	2	11	5	0	6	10	5	0
	3	0	0	1	0	0	0	1	0

Tabla 15. Matriz de confusión para los rasgos G y R con parametrización de los índices de ruido y CEM de 26 centroides (100 ms).

Con la combinación de parámetros, los resultados de validación cruzada arrojan una eficiencia del **53 %** (varianza: 8 %) para el rasgo **G** y del **51 %** (varianza: 6 %) para **R**. Entrenamientos realizados con otras variables de parametrización no proporcionan mejores resultados.

7. Discusión y trabajo futuro

Como puede observarse en la Tabla 16, los resultados obtenidos con los centroides, a pesar de su simplicidad, son ligeramente superiores a los que proporcionan los coeficientes MFCC para los dos rasgos estudiados. También son superiores a los proporcionados por la parametrización del ruido para el rasgo G. A la vista de los resultados se puede afirmar que con la base de datos utilizada el método de parametrización más efectivo es el de los índices de complejidad.

	Correlación cruzada. Eficiencia media % (varianza) %						
Rasgo	CEM	MFCC	MFCC + CEM	Complejidad	Complejidad + CEM	Ruido	Ruido + CEM
G	59 (7)	51 (7)	55 (6)	60 (6)	59 (7)	55 (10)	53 (8)
R	48 (5)	48 (8)	49 (6)	53 (6)	53 (5)	49 (7)	51 (6)

Tabla 16. Resumen de los resultados de validación cruzada obtenidos para las distintas parametrizaciones usando tramas de 100 ms y solapamiento del 50%.

Sin embargo, y debido a que la información proporcionada por los centroides acerca del espectro de modulación es muy simple, cabe esperar que una extracción de características del EM más completa produzca mejores resultados. Un posible nuevo parámetro basado en el EM sería la relación entre la energía entorno a 0 Hz y el resto de energía hasta 25 Hz para las primeras bandas acústicas (normalmente en torno a la frecuencia fundamental). Se propone esta medida ya que se observa que la relación entre la energía a 0 Hz y el resto de frecuencias de modulación es siempre menor en voces patológicas. Este podría ser un primer parámetro a evaluar. También existe la posibilidad de medir los valores de las frecuencias de modulación más representativas en la banda de la frecuencia fundamental y el ancho de banda de modulación para cada una de ellas. La tercera posibilidad que se propone consiste en calcular la densidad espectral de potencia (Power Spectral Density - PSD) mediante el método Welch y calcular su desviación Estándar en la banda que incluya todas las frecuencias

de interés. Esto nos daría una cierta información sobre los cambios en la dinámica del EM.

Además, para conocer mejor cómo se ven representados en el EM los síntomas provocados por algunas enfermedades, como puede ser el tremor, distintos tipos de ronquera o shimmer entre muchos otros, puede ser recomendable usar un vocoder para generar señales conocidas y observar el comportamiento de su EM. A partir de esto, podríamos sacar algunas conclusiones más fácilmente y aplicarlas a la interpretación del EM a señales de voz reales.

En cuanto a la fusión de características, en el caso del rasgo G, se observa que la única mejora se produce con la fusión CEM-MFCC en donde se pasa de una eficiencia media del 51 % (sólo MFCC) al 55% (CEM-MFCC). Sin embargo, desde otro punto de vista, dicha fusión no es del todo provechosa ya que pasaríamos de una eficiencia media del 59 % con parámetros CEM a 55 % con la fusión MFCC-CEM. Es por esto que dentro del marco de esta investigación sería más aconsejable un clasificador con parámetros CEM que uno con fusión MFCC-CEM. Es destacable el hecho de que en todas las fusiones para el rasgo G disminuye la eficiencia media con respecto al uso de una única familia de parámetros. Posiblemente esto se deba a que por un lado la complementariedad de las características no es lo suficientemente relevante. Por otro lado, el aumento en el número de parámetros complica el espacio multidimensional de los sistemas de clasificación SVM empobreciéndose la eficiencia media de clasificación. Muy probablemente esta segunda causa sea la que tiene más peso. Por lo tanto, sería recomendable repetir en el futuro las pruebas realizadas utilizando algún método de selección de las características más importantes, como el de máxima relevancia [54].

Algo distinto ocurre con el rasgo R. En todos los casos la fusión de características iguala o aumenta la precisión media. Estos resultados sugieren que para R la cantidad de información complementaria de unas características a otras sí es suficientemente relevante como para dar mejores resultados. Todo ello a pesar de que el aumento en el número de parámetros de entrada en el clasificador SVM complique el espacio de multidimensional. Igualmente, el uso de selección de características de máxima relevancia podría aumentar la eficiencia. No obstante se debe tener en cuenta que la utilización de las técnicas de selección de características puede llevar en determinados casos a obtener modelos demasiado ajustados a la base de datos utilizada, perdiéndose generalización. En estos casos se haría necesaria una base de datos lo suficientemente extensa y heterogénea y una justificación clara de los motivos que nos llevan a seleccionar unas características frente a otras.

Continuando con el análisis de los resultados, al observar las matrices de confusión se comprueba que por norma general todos los errores de clasificación cometidos recaen en niveles adyacentes al nivel supuestamente real. Este comportamiento es bastante coherente con el hecho de que las etiquetas utilizadas provienen de una evaluación subjetiva. Dicha subjetividad conlleva una cierta falta de consistencia en el etiquetado, que podría hacer que dos registros de audio con características acústicas muy similares sean etiquetados, por parte de un mismo evaluador, con niveles cercanos pero no idénticos.

En estas matrices también puede observarse que en los dos últimos niveles (especialmente en el nivel con valor 3) la eficiencia baja con respecto a los dos primeros. Esto es debido al desbalance de clases presente en la base de datos. La solución tomada en este proyecto ha sido descartar tramas de las clases con mayor representación durante el entrenamiento. En realidad sería más aconsejable ampliar la base de datos existente para incluir un número mayor de voces patológicas que estén dentro de estos niveles. Si esto no fuese posible, en trabajos futuros se debería plantear el uso de algoritmos de balance de clases [84].

A modo de ejemplo, a continuación se muestra la variabilidad inter-evaluador e intra-evaluador del etiquetado del rasgo G de la base de datos de Kay Elemetrics [13] realizado por dos evaluadores distintos. Uno de ellos realizó el etiquetado dos veces y el otro sólo una. Se exponen estos resultados en las Tablas 17 y 18 como muestra de un caso conocido de variabilidad sin implicar necesariamente que los valores obtenidos sean representativos del caso general.

		Evaluación 2			
Evaluación 1	Clases	0	1	2	3
	0	78	5	0	0
	1	1	21	0	0
	2	0	19	18	0
	3	0	0	20	59

Tabla 17. Matriz de confusión de dos evaluaciones del Evaluador 1 en dos momentos distintos respecto a la misma base de datos.

En este caso podemos considerar que la eficiencia intra-evaluador es del 79,6%

	Evaluador 2				
Evaluador 1	Clases	0	1	2	3
	0	30	44	9	0
	1	2	7	13	0
	2	1	8	26	2
	3	0	10	30	39

Tabla 18. Matriz de confusión del Evaluador 1 respecto al Evaluador 2

Ahora la eficiencia inter-evaluador disminuye al 46%.

Así pues, aun cuando los resultados de eficacia media obtenidos en este proyecto pudieran parecer comedidos, se considera que los sistemas de clasificación obtenidos se aproximan bastante al caso real.

Cabe destacar que en el presente proyecto se busca analizar el comportamiento de los sistemas de clasificación de niveles perceptuales frente a distintas parametrizaciones. Es por eso que se dan los resultados obtenidos para los dos rasgos G y R utilizando las mismas características y longitudes de trama. En el desarrollo de un detector de cada uno de los rasgos del GRBAS, deberían estudiarse todas las parametrizaciones y entrenamientos posibles por separado, pudiéndose obtener longitudes de trama y número de parámetros distintos para cada uno de ellos. Esto es así debido a que cada rasgo está referido a propiedades de la voz distintas y requiere un tratamiento distinto.

En cuanto a la metodología, estudios como [42] o [43] apuntan a que el uso de una base de datos en la que exista un solo género (masculino o femenino) podría proporcionar mejores resultados. En [42] se llega a obtener una eficiencia del 85 % si bien no se puede establecer una comparación consistente con este proyecto ya que se utilizan bases de datos y evaluadores distintos. Igualmente, sería recomendable el estudio de una posible segmentación de las bases de datos por edades de los locutores además de por género, lo que podría tener consecuencias positivas.

Por otro lado, se ha de tener en cuenta que la evaluación GRBAS se realiza mediante la escucha de una o más vocales sostenidas y de habla continua para poder así comprobar diversas características necesarias en la voz [1]. De este modo podría ser relevante añadir la parametrización del habla continua a la realizada en este proyecto porque proporciona información que es tenida en cuenta en la valoración perceptual subjetiva.

En cuanto a los tiempos de cómputo utilizados para obtener cada uno de los parámetros, la Tabla 19 muestra la relación entre los parámetros calculados.

Parámetros	CEM	MFCC	Complejidad	Ruido
Tiempo de cómputo	7h 14 min	2 min	2h 22min	11min

Tabla 19. Tiempos de cómputo de los parámetros para el corpus reducido usando tramas de 50 ms y solapamiento del 50%.

Claramente, el proceso de cálculo de CEM tiene un tiempo de cómputo muy superior al resto, lo que sugiere una optimización de los algoritmos para su uso con grandes bases de datos o en aplicaciones que requiriesen procesado en tiempo real.

8. Conclusiones

Los nuevos parámetros CEM proporcionan información representativa sobre el nivel de afección presente en voces patológicas. El uso de CEM arroja valores de eficiencia algo superiores al uso de MFCC en simulaciones de evaluación perceptual automática y muy similares a los de las parametrizaciones basadas en complejidad. Una fusión de características entre los nuevos parámetros CEM y otros utilizados anteriormente aumenta en algunos casos la efectividad media pero se recomienda el estudio de técnicas de reducción de características en trabajos futuros.

Igualmente, se demuestra que los sistemas de clasificación SVM pueden ser válidos para la simulación de evaluaciones perceptuales. Estos clasificadores habían sido utilizados anteriormente en trabajos similares pero no en evaluación perceptual.

Los siguientes pasos para continuar con esta investigación consistirán en el desarrollo de nuevos parámetros relacionados con el EM y la extracción de las características más relevantes.

Referencias

[1] M. Hirano, *Clinical examination of voice*. Springer Verlag, 1981.

[2] C. Sapienza and B. Hoffman Ruddy, *Voice Disorders*. Plural Publishing, 2009.

[3] A. L. Webb, P. N. Carding, I. J. Deary, K. MacKenzie, N. Steen, and J. A. Wilson, "The reliability of three perceptual evaluation scales for dysphonia.," *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, vol. 261, no. 8, pp. 429–34, Sep. 2004.

[4] R. Schönweiler, M. Hess, P. Wübbelt, and M. Ptok, "Novel approach to acoustical voice analysis using artificial neural networks," *JARO-Journal of the Association for Research in Otolaryngology-*, vol. 1, no. 4, pp. 270–282, 2000.

[5] G. B. Kempster, B. R. Gerratt, K. Verdolini Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol.," *American journal of speech-language pathology / American Speech-Language-Hearing Association*, vol. 18, no. 2, pp. 124–32, May 2009.

[6] I. V. Bele, "Reliability in perceptual analysis of voice quality.," *Journal of voice : official journal of the Voice Foundation*, vol. 19, no. 4, pp. 555–73, Dec. 2005.

[7] M. S. De Bodt, F. L. Wuyts, P. H. Van de Heyning, and C. Croux, "Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality.," *Journal of voice : official journal of the Voice Foundation*, vol. 11, no. 1, pp. 74–80, Mar. 1997.

[8] P. H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman, and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements.," *Revue de laryngologie - otologie - rhinologie*, vol. 117, no. 3, pp. 219–24, Jan. 1996.

[9] I. Titze and D. Martin, "Principles of voice production," *The Journal of the Acoustical Society of America*, 1998.

[10] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on acoustics, Speech and Signal processing*, vol. ASP-27. NO, 1979.

[11] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, 2000.

[12] D. Povey, S. M. Chu, and B. Varadarajan, "Universal background model based speech recognition," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, no. 5, pp. 4561–4564, Mar. 2008.

[13] "Voice Disorders Database." Massachusetts Eye and Ear Infirmary, 1994.

[14] M. Liu, B. Dai, Y. Xie, and Z. Yao, "Improved GMM-UBM/SVM For Speaker Verification," *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, vol. 1, pp. I–925–I–928, 2006.

[15] M. Markaki, S. Member, and Y. Stylianou, "Voice Pathology Detection and Discrimination Based on Modulation Spectral Features," vol. 19, no. 7, pp. 1938–1948, 2011.

[16] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices.," *Logopedics, phoniatrics, vocology*, vol. 36, no. 2, pp. 60–9, Jul. 2011.

[17] M. a Little, P. E. McSharry, S. J. Roberts, D. a E. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection.," *Biomedical engineering online*, vol. 6, p. 23, Jan. 2007.

[18] W. Winholtz and L. Ramig, "Vocal tremor analysis with the vocal demodulator," *Journal of Speech, Language and Hearing Research*, 1992.

[19] J. Godino-Llorente, "Discriminative methods for the detection of voice disorders," *NOLISP-2005*, pp. 158–167, 2005.

[20] J. I. Godino-Llorente and P. Gómez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE transactions on bio-medical engineering*, vol. 51. NO2, 2004.

[21] J. J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement.," *Journal of voice : official journal of the Voice Foundation*, vol. 20, no. 1, pp. 2–17, Mar. 2006.

[22] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[23] M. Marinaki and C. Kotropoulos, "Automatic detection of vocal fold paralysis and edema," *Proceedings of ICSLP '04, Jeju Island, South Korea*, 2004.

[24] K. Shama, A. Krishna, and N. U. Cholayya, "Study of Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 085286, Jan. 2007.

[25] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders.," *Journal of voice : official journal of the Voice Foundation*, vol. 24, no. 1, pp. 47–56, Jan. 2010.

[26] J. Alonso, J. de León, I. Alonso, and M. A. Ferrer, "Automatic detection of pathologies in voice by HOS based parameters," *Journal on Applied Signal Processing. EURASIP*, 2001.

[27] P. Gómez-Vilda, R. Fernández-Baillo, and A. Nieto, "Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters," *Journal of Voice*, 2007.

[28] B. Ghoraani and S. Krishnan, "A joint time-frequency and matrix decomposition feature extraction methodology for pathological voice classification," *EURASIP Journal on Advances in Signal Processing*, 2009.

[29] A. A. Dibazar, S. Narayanad, and T. W. Berger, "Feature Analysis for Automatic Detection of Pathological Speech," *Proceedings of the Second Joint EMBS/BMES Conference, Houston, TX, USA.*, pp. 0–1, 2002.

[30] M. Wester, "Automatic classification of voice quality: Comparing regression models and hidden markov models," *Proceedings of Voicedata '98, Utretch, The Netherlands*, pp. 92–97, 1998.

[31] L. Leinonen, J. Kangas, K. Torkkola, and A. Juvas, "Dysphonia detected by pattern recognition of spectral composition," *Journal of Speech, Language and Hearing Research*, 1992.

[32] G. Schlotthauer, M. Torres, and H. Rufiner, "Pathological voice analysis and classification based on empirical mode decomposition," *Development of multimodal interfaces: Active Listening and Synchrony*, vol. 5967, p. 364–, 2010.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, 1995.

[34] V. N. Vapnik, "An overview of statistical learning theory.," *IEEE transactions on neural networks/ IEEE Neural Networks Council*, vol. 10, no. 5, pp. 988–99, Jan. 1999.

[35] P. Kukharchik and I. Kheidorov, "Speech signal processing based on wavelets and SVM for vocal tract pathology detection," *Image and Signal Processing. Springer Berlin Heidelberg.*, pp. 192–199, 2008.

[36] P. Taylor, B. Efron, and G. Gong, "A Leisurely Look at the Bootstrap , the Jackknife , and A Leisurely Look at the Bootstrap , the Jackknife , and," *The American Statician*, no. April 2013, pp. 37–41, 2012.

[37] H. Kasuya, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, p. 1329, Nov. 1986.

[38] A. Schuck Jr, L. V. Guimaraes, and J. O. & Wisbeck, "Dysphonic voice classification using wavelet packet transform and artificial neural network," *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 3, pp. 2958–2961, 2003.

[39] K. Umapathy and S. Krishnan, "Discrimination of pathological voices using a time-frequency approach," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2002.

[40] B. Ghoraani, K. Umapathy, L. Sugavaneswaran, and S. Krishnan, "Pathological speech signal analysis using time-frequency approaches.," *Critical reviews in biomedical engineering*, vol. 40, no. 1, pp. 63–95, Jan. 2012.

[41] J. I. Godino-Llorente, P. Gómez-Vilda, F. Cruz-Roldán, M. Blanco-Velasco, and R. Fraile, "Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness.," *Journal of voice : official journal of the Voice Foundation*, vol. 24, no. 6, pp. 667–77, Nov. 2010.

[42] G. Pouchoulin, C. Fredouille, J. Bonastre, A. Ghio, and A. Giovanni, "Dysphonic Voices and the 0-3000Hz Frequency Band," *Interspeech 2008. ISCA*, pp. 2214–2217, 2008.

[43] G. Pouchoulin, C. Fredouille, J. Bonastre, A. Ghio, A. Giovanni, A. France, P. France, and M. France, "Frequency Study for the Characterization of the Dysphonic Voices," *Interspeech 2007. ISCA*, pp. 1198–1201, 2007.

[44] G. Pouchoulin, C. Fredouille, J. Bonastre, A. Ghio, and J. Revis, "Characterization of the pathological voices (dysphonia) in the frequency space," *Proceedings of International Congress of Phonetic Sciences (ICPhS)*, no. August, pp. 1993–1996, 2007.

[45] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the GRBAS scale.," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society.*, vol. 1, pp. 2478–81, Jan. 2006.

[46] G. Toussaint, "Bibliography on estimation of misclassification," *IEEE Transactions on Information Theory*, 1974.

[47] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, J. M. Gutiérrez-Arriola, and V. Osma-Ruiz, "Spectral analysis of pathological voices: sustained vowels vs running speech," *Models and Analysis of Vocal Emissions for Biomedical Applications 7th international Workshop. Firenze University Press.*, vol. 77, p. 67, 2011.

[48] S. Schimmel, L. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," *EEE International Conference in Acoustics, Speech and Signal Processing, 2007. ICASSP*, vol. 4, 2007.

[49] T. F. Q. Nicolas Malyska, "Automatic dysphonia recognition using biologically inspired amplitude-modulation features," *Proc. ICASSP*, vol. 1, pp. 873–876.

[50] M. Markaki and Y. Stylianou, "Modulation Spectral Features for Objective Voice Quality Assessment: The Breathiness Case," *Sixth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications.*, 2009.

[51] J. I. Markaki, M., Stylianou, Y., Arias-Londono, J. D., & Godino-Llorente, "Dysphonia detection based on modulation spectral features and cepstral coefficients," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 5162–5165, 2010.

[52] M. Markaki; Holzapfel, Andre; Stylianou, Yannis. "Singing voice detection using modulation frequency features". *Workshop on Statistical and Perceptual Audition*. 2008. p. 7-10. , 2010.

[53] M. Markaki and Y. Stylianou, "Modulation Spectral Features for Objective Voice Quality Assessment: The Breathiness Case," *Sixth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications.*, 2009.

[54] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1226–1238, 2005.

[55] R. Shrivastav, "The use of an auditory model in predicting perceptual ratings of breathy voice quality," *Journal of Voice*, vol. 17, no. 4, pp. 502–512, Dec. 2003.

[56] R. Shrivastav and C. M. Sapienza, "Some difference limens for the perception of breathiness," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, p. 416, Jul. 2006.

[57] S. Anand, R. Shrivastav, J. M. Wingate, and N. N. Chheda, "An acoustic-perceptual study of vocal tremor.," *Journal of voice : official journal of the Voice Foundation*, vol. 26, no. 6, pp. 811.e1–7, Nov. 2012.

[58] J. Gamboa, F. J. Jiménez-Jiménez, a Nieto, I. Cobeta, a Vegas, M. Ortí-Pareja, T. Gasalla, J. a Molina, and E. García-Albea, "Acoustic voice analysis in patients with essential tremor.," *Journal of voice : official journal of the Voice Foundation*, vol. 12, no. 4, pp. 444–52, Dec. 1998.

[59] "Les Atlas, Pascal Clark and Steven Schimmel, Modulation Toolbox Version 2.1 for MATLAB, http://isdl.ee.washington.edu/projects/modulationtoolbox/, University of Washington, September 2010." .

[60] B. Gajic and K. K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 600–608, Mar. 2006.

[61] G. Arias-Londono, J. D., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., & Castellanos-Dominguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, pp. 370–379, 2011.

[62] Y. Zhang, J. J. Jiang, L. Biazzo, and M. Jorgensen, "Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis.," *Journal of voice : official journal of the Voice Foundation*, vol. 19, no. 4, pp. 519–28, Dec. 2005.

[63] A. Giovanni, M. Ouaknine, and J. Triglia, "Determination of largest Lyapunov exponents of vocal signal: application to unilateral laryngeal paralysis," *Journal of Voice*, 1999.

[64] H. Kantz and T. Schreiber, "Nonlinear time series analysis," *Cambridge University Press*, vol. 7, 2003.

[65] S. Pincus, "Approximate entropy as a measure of system complexity," *Proceedings of the National Academy of Sciences*, vol. 88(7), pp. 2297–2301, 1991.

[66] I. Rezek and S. Roberts, "Stochastic complexity measures for physiological signal analysis," *Transactions on Biomedical Engineering, IEEE.*, vol. 45 (9), pp. 1186–1191, 1998.

[67] J. Richman and J. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circualtory Physiology*, vol. 278(6), pp. H2039–H2049, 2000.

[68] H.-B. Xie, W.-X. He, and H. Liu, "Measuring time series regularity using nonlinear similarity-based sample entropy," *Physics Letters A*, vol. 372, no. 48, pp. 7140–7146, Dec. 2008.

[69] L. Xu, K. Wang, and L. Wang, "Gaussian kernel approximate entropy algorithm for analyzing irregularity of time-series," *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, vol. 9, pp. 5605–5608, 2005.

[70] B. Kosko, "Fuzzy entropy and conditioning," *Information sciences*, 1986.

[71] M. Little, D. Costello, and M. Harries, "Objective dysphonia quantification in vocal fold paralysis: comparing nonlinear with classical measures," *Journal of Voice*, 2011.

[72] P. Yu, J. Revis, F. L. Wuyts, M. Zanaret, and A. Giovanni, "Correlation of Instrumental Voice Evaluation with Perceptual Voice Analysis Using a Modified Visual Analog Scale," *Folia Phoniatr Logop*, 2002.

[73] A. Schindler, F. Palonta, G. Preti, F. Ottaviani, O. Schindler, and A. L. Cavalot, "Voice quality after carbon dioxide laser and conventional surgery for T1A glottic carcinoma.," *Journal of voice : official journal of the Voice Foundation*, vol. 18, no. 4, pp. 545–50, Dec. 2004.

[74] V. Parsa and D. G. Jamieson, "Identification of Pathological Voices Using Glottal Noise Measures," *J Speech Lang Hear Res*, vol. 43, no. 2, pp. 469–485, Apr. 2000.

[75] E. Yumoto, "The quantitative evaluation of hoarseness: A new harmonics to noise ratio method," *Archives of Otolaryngology—Head & Neck Surgery*, 1983.

[76] D. Michaelis, "Glottal-to-noise excitation ratio a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[77] D. Deliyski, "Acoustic model and evaluation of pathological voice production," *Proceedings of Eurospeech*, 1993.

[78] G. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language and Hearing Research*, 1993.

[79] G. Jotz, O. Cervantes, and M. Abrahão, "Noise-to-harmonics ratio as an acoustic measure of voice disorders in boys," *Journal of voice*, 2002.

[80] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," *National Taiwan University*, vol. 1, no. 1, pp. 1–16, 2010.

[81] "Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm." .

[82] P. Taylor, G. H. Golub, M. Heath, and G. Wahba, "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, vol. 21(2), no. April 2013, pp. 215–223, 1979.

[83] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, and C. Ramírez-Calvo, "Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program.," *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, vol. 265, no. 4, pp. 465–76, Apr. 2008.

[84] E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.