Universidad Zaragoza
1542

Trabajo Fin de Grado

Análisis y Detección de Edema de Reinke mediante Procesado de Señal y Aprendizaje Automático

Autor/es

Daniel Quintilla Ruipérez

Director/es

Eduardo Lleida Solano

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

Escuela de Ingeniería y Arquitectura

2020

Agradecimientos

A mi tutor Eduardo Lleida por su asesoramiento y apoyo, siempre desde una posición amable y paciente.

A mi madre, mi padre y mi hermana por su cariño incondicional.

A María, por ser mi mayor apoyo durante todo este tiempo.

A mis amigos y amigas de siempre, que siempre me han animado en momentos difíciles.

Y en especial a los amigos que he conocido en estos 4 años de universidad sin los cuales no habría sido posible llegar hasta aquí.

"No nacimos para resistir, nacimos para vencer"

-Ricardo Romero

Índice general

Resumen

Esta memoria describe el desarrollo de un sistema de análisis y detección de la disfonía conocida como Edema de Reinke a través de técnicas de procesado digital de señal y algoritmos de aprendizaje automático tales como las Redes Neuronales.

Además, se comentarán los estudios y desarrollos similares realizados anteriormente en los que me he basado y de los que he aprendido para llevar a cabo mi trabajo.

La intención de este trabajo es la de crear una base sobre la que construir un proyecto de mayor escala centrado en la detección de múltiples enfermedades que afectan a la voz mediante métodos no invasivos, acompañado de la creación de una base de datos de gran escala, en la que se incluyan grabaciones de voz de pacientes con afecciones vocales tanto antes como después de la operación o tratamiento necesario y encuestas sobre hábitos de estos pacientes.

Este proyecto comenzó hace unos meses con el nombre de THALENTO (Tecnologías del HAbla y el Lenguaje para la EvaluaciónN de Transtornos de la cOmunicación) pero fue interrumpido temporalmente por el COVID-19 por lo que el estudio estadístico de ciertos parámetros vocales antes y después del tratamiento han sido realizados solamente sobre 3 sujetos, por lo que constituye una mínima parte del trabajo.

La carga principal del trabajo se centra en el sistema de detección del edema de Reinke en las cuerdas vocales mediante algoritmos de aprendizaje automático, cuyo desarrollo consta de los siguientes pasos: La extracción de numerosos parámetros utilizados habitualmente para estudios sobre afecciones de la voz, su correcta organización, etiquetado y almacenaje separándolos en distintos sets de datos; y el diseño de varios modelos de redes neuronales, uno específico para cada grupo de datos. De esta manera podremos analizar que grupos de datos contienen una mayor cantidad de información y son más relevantes en este aspecto y así poder diseñar en un futuro sistemas más eficientes que necesiten menos información para obtener los mismos resultados.

Finalmente, se llevará a cabo un estudio de la eficiencia de los diferentes sistemas de clasificación para analizar que algoritmos se adaptan mejor a los datos que tenemos y por qué, y para comprobar finalmente si es posible tener porcentajes de acierto suficiente como para considerar el sistema lo suficientemente útil como para usarse clínicamente.

Abstract

This report describes the development of a system of analysis and detection of the dysphonia known as Reinke's Edema through Signal Digital Processing techniques and machine learning algorithms such as Neural Networks.

In addition, some of the last studies in this área of knowledge, from where this job has been based, will be commented.

The aim of this work is to create a baseline on which to build a larger scale proyect centered in the detecction of several voice diseases through non-invasive methods, accompanied by the creation of a large-scale database, which will include voice recordings of patients with vocal affections both after and before of the surgery or medical intervention and surveys about the habits of the patients.

This project started some months ago with the name of THALENTHO but unfortunately was interrupted by the COVID-19 so we count only with recordings from both after and before of 3 patients.

The main part of this job is focused on a system of Reinke's Edema detection using machine learning algorithms, whom development consists of the following steps: The extraction of several parameters commonly used in studies about voice diseases and its correct organization, labelled and storage; and the design of some systems of machine learning.

Finally, a study of accuracy of the different systems will me presented, analyzing which algorithms fit better to the data and why, and for considering if the system is reliable and useful enough for using it medically.

Índice de figuras

Índice de Tablas

Acrónimos

ER: Edema de reinke

VHI: Voice Handicap Index

TMF: Tiempo máximode Fonación

SVD: Saarbruecken Voice Database

ECV: Encuesta del Comportamiento Vocal

NHR: Noise to Harmonic Ratio

HNR: Harmonic to Noise Ratio

VTI: Voice Turbulence Index

SPI: Soft Phonation Index

GNE: Glottal to Noise Excitation

NNE: Normalized Noise Energy

MFCC: Mel Frequency Cepstrum Coefficients

DCT: Discrete Cosine Transform

CPP: Cepstral Peak Prominence

FFT: Fast Fourier Transform

MNN: Multilayer Neural Network

1. Introducción y objetivos

1.1. Contexto

Actualmente, y teniendo en cuenta el ritmo de evolución tecnológica de los últimos años, la inteligencia artificial está cada vez más presente en nuestras vidas cotidianas; por ejemplo, muchos de nosotros le pedimos a Alexa que nos ponga música o a Siri que nos recuerde que compremos pan, tenemos cámaras que nos detectan la cara en las fotos, o Spotify y Netflix nos recomiendan canciones y películas a partir de lo que hemos escuchado o visto anteriormente.

Los algoritmos de aprendizaje automático han revolucionado el panorama tecnológico con sus altas eficiencias en aplicaciones de clasificación, predicción o detección, por eso cada vez más, se utilizan dichos algoritmos para hacer diagnósticos médicos prematuros. Gracias a la cantidad de información que aportan nuestras señales vitales y a las técnicas de reconocimiento de imágenes muchas patologías pueden ser detectadas sin necesidad de acudir al médico, algo muy útil en épocas como la vivida recientemente. De hecho, durante esta etapa se ha demostrado que los sistemas de telemedicina son muy útiles y permiten agilizar muchos trámites. En Estados Unidos, se realizaron más de 2 millones de tele consultas en 2015, y en 2019 la compañía UnitedHealth lanzó una aplicación de atención médica virtual para más de 27 millones de personas (Hospital, 2020).

Dichos sistemas son necesarios en nuestro país, espacialmente en ambientes rurales y con pacientes crónicos y de avanzada edad. Evitando así, desplazamientos innecesarios que, en el caso de algunas zonas rurales, pueden tratarse de decenas de kilómetros. (Healthtech, 2020)

Esto puede hacer que alguien llegue a la conclusión de que las máquinas sustituirían a los médicos, pero en absoluto es así; estos algoritmos ofrecen finalmente un conjunto de probabilidades que pueden orientar de alguna manera a pacientes u ofrecer más información al médico en el diagnóstico.

Otro aspecto a tener en cuenta y que fomenta la investigación de este tipo de desarrollos es que las técnicas de diagnóstico para desordenes en la voz son muy invasivas; pruebas como la laringoscopia son más complicadas y suponen al paciente un esfuerzo mayor que el que supone un análisis acústico de su voz y así quizás evitar

pruebas de este tipo. Además, esta clase de patologías son más comunes de lo que creemos, según la Sociedad Española de Otorrinolaringología (Mateos & Jiménez, 2016) "Uno de cada 13 personas sufre trastornos de la voz, pero la mayoría no se tratan adecuadamente". En este porcentaje los grupos que se ven más afectados son el de jóvenes y los profesores.

1.2. Estudios similares

En este apartado se comentan algunos de los estudios más recientes en el campo de la detección de disfonías a partir de parámetros vocales.

En el artículo "Voice Pathology Detection Using Deep Learning: a Preliminary Study" (Harar et al, 2017) publicado en el IEEE se explica el desarrollo del sistema en el que se extraen muestras de voz del SVD, únicamente de la vocal 'a'.

Se utilizan fragmentos de 64 ms muestreados a 50kHz como entrada a una Red Neuronal Profunda con 11 capas en las que se disminuyen las dimensiones de los datos, con la finalidad de clasificar entre voz patológica o voz sana. Se consiguen precisiones de entre 68,08% y 71.36% con el set de test de datos. Nos encontramos con un sistema en el que el procesado de los datos no es complejo y que centra su esfuerzo en el desarrollo de la red neuronal; pero como podemos observar, las precisiones finales no son muy altas comparadas con las que se obtienen en otros sistemas que comentaremos más adelante.

En el siguiente artículo (Hegde et al, 2018) se hace un recorrido por los distintos algoritmos de aprendizaje automático tales como los Hidden Markov Models (HMM), Support Vector Machines (SVM), Artificial Neural Networks (ANN), arboles de decisión, agrupamiento por K-medias o clasificadores combinados, para evaluar su eficiencia en la detección de desórdenes de la voz. Como entrada a estos clasificadores se utilizaron parámetros como los Coeficientes de Predicción Lineal (LPC), Mel-frequency Cepstral Coefficient (MFCC), parámetros de flujo glotal y parámetros acústicos, muchos de ellos han sido utilizados en este desarrollo por lo que serán explicados en profundidad próximamente.

Otro artículo a comentar, y uno de los que ha resultado esencial para este trabajo es el titulado "Vocal Acoustic Analysis - Classification of Dysphonic Voices with Artificial Neural Networks" (Teixeira et al, 2017). El cual explica de manera explícita los parámetros que ha utilizado en su desarrollo, como, por ejemplo, algunos muy importantes como el Jitter, Shimmer o HNR.

Estos estudios comentados anteriormente siguen un patrón similar que consta de un procesado de muestras de voz para extraer la información deseada, la selección del tipo de clasificador más apropiado y su correspondiente diseño en base a la estructura de los datos procesados.

1.3. Edema de Reinke

El Edema de Reinke es un proceso inflamatorio de las cuerdas vocales en el que se acumula liquido entre el musculo vocal y la mucosa que lo cubre.

1.3.1. Causas

Esta afección mayoritariamente ocurre como respuesta a un consumo prolongado del tabaco, aunque también se puede dar por un abuso de la voz como sería el caso de profesores, cantantes u operadores; o también por culpa del reflujo faringolaríngeo. (Reyes Burneo, 2018). Generalmente esta patología es bilateral, en algunas ocasiones, solo afecta a una cuerda vocal, afectando en mayor medida a mujeres que a hombres. En la SVD encontramos 657 muestras de voz de mujeres afectadas por Edema de Reinke mientras que solo contamos con 72 muestras de hombres afectados. Según (Reyes Burneo) la proporción es de 5:1 aunque con los datos de la SVD podemos afirmar que es incluso mayor.

1.3.2. Síntomas

El síntoma principal es una intensa alteración de la voz fácilmente perceptible por el oído humano, agravando el tono de voz de la persona en cuestión. También puede causar tos y carraspera y en casos graves puede derivar en dificultades respiratorias. (ISEP, 2017)

1.3.3. Diagnóstico

Cuando aparece un paciente con los síntomas se debe investigar su hábito con el tabaco, la cantidad que consume y durante cuánto tiempo, también se debe tener en cuenta

el uso habitual que le da a la voz y finalmente se le realiza una exploración que incluye (A. Gonzalez et al, 2020):

- Valoración objetiva:
  - Laringoscopia indirecta
  - Análisis acústico
  - Estudio aerodinámico
- Valoración subjetiva:
  - Escala GRABS
  - Índice de incapacidad vocal (VHI)

Durante el diagnóstico es esencial poder diferencial el Edema de Reinke con el resto de las denominadas “lesiones exudativas del espacio de Reinke”, que son:

- Pólipo vocal
- Nódulo vocal
- Pseudoquiste seroso

Dependiendo del tamaño del edema y el nivel en que se encuentre se puede clasificar de acuerdo con la clasificación de Yonekawa (A. Gonzalez et al, 2020):

- Grado I: El edema se encuentra solamente en la parte superior de las cuerdas vocales.
- Grado II: El edema está más extendido, uniendo ligeramente las dos cuerdas vocales.
- Grado III: El edema es tan extenso que puede llegar a impedir hasta dos tercios del paso aéreo.

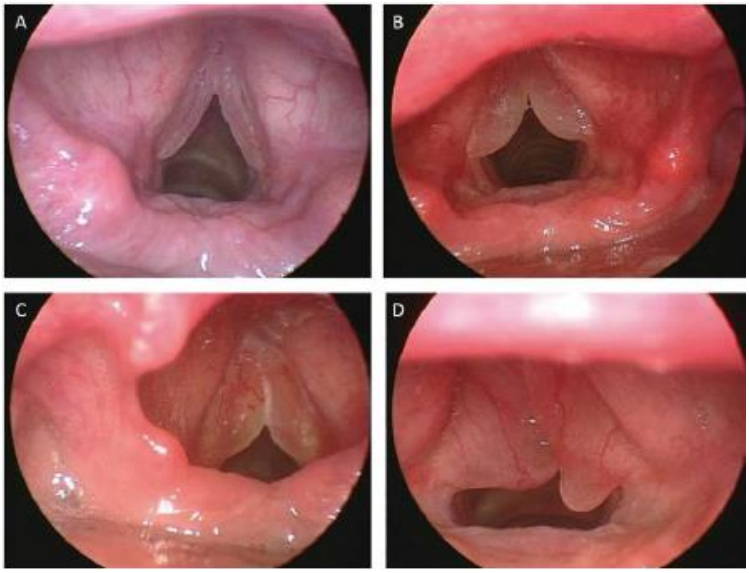

Figura 1. Imagen endoscópica con visión directa de ER. A: grado I, B: grado II, C: grado III, D: grado III con formación polipoidea, según clasificación de Yonekawa.

1.3.3.1. Laringoscopia indirecta

Se basa en introducir un fibroscopio flexible por una fosa nasal hasta llegar a los pliegues vocales en los que se pueden apreciar el movimiento de las cuerdas vocales durante el habla. Si dicha técnica se realiza con luz estroboscópica se denomina laringoestroboscopia y gracias a ella se pueden valorar factores como:

- Cierre glótico.
- Frecuencia de pitch
- Simetría.
- Amplitud
- Onda Mucosa

1.3.3.2. Estudio aerodinámico

Aquí se analizan parámetros relacionados con las presiones y flujos aéreos que participan en la fonación (A. Gonzalez et al), los más importantes son:

- Tiempo máximo de fonación (TMF): Mide el tiempo máximo que el paciente puede aguantar sosteniendo una vocal, deben hacerlo 3 veces y se guarda el máximo valor.

- Índice fonorespiratorio: Se trata de obtener el cociente entre el TMF realizado con el fonema /s/ y el fonema /a/ para comprobar la diferencia entre un fonema que depende de las cuerdas vocales frente a otro que depende únicamente de la capacidad pulmonar.

1.3.3.3. Escala GRABS

Su nombre se debe a las descripciones que trata de evaluar.

- Grade (Grado general): Se refiere al grado de ronquera en la voz.
- Roughness (aspereza): Referida a la sequedad de la voz y su falta de proyección.
- Asteny (astenia): Energía en la emisión de la voz.
- Breathiness (escape aéreo): Referido a como de entrecortada se escucha la voz.
- Strain (Tension): Opresión en las cuerdas vocales.

Cada descripción se evalúa de 0(normal) a 3(severo)

1.3.3.4. VHI

El VHI o su traducción al español "Índice de Incapacidad Vocal", es una encuesta de 30 preguntas sobre el impacto que generan dicha disfonía en su vida diaria. Se puntúan de 0 a 4. Tras sumar las puntuaciones de todas las respuestas, cuanto más alta sea la suma, más incapacitante es la disfonía. (ANEXO II: Encuestas VH).

1.3.4. Tratamiento

En la mayoría de los casos, que a su vez se diagnostican como leves, el tratamiento se basa únicamente en suprimir el elemento causante de la enfermedad, siendo el principal agente el tabaco. Si es cierto que, aunque el edema desaparezca con la supresión del tabaco, habitualmente no es suficiente para recuperar la normalidad en la voz.

Existen ciertos tratamientos médicos y logopédicos que junto a la supresión del tabaco pueden repercutir ciertas mejoras, pero siguen sin ser suficiente para mejorar completamente la calidad de la voz.

El tratamiento más efectivo y que muestra unas mejoras muy considerables en un espacio de tiempo muy corto es el tratamiento quirúrgico, que se basa en acceder al espacio de Reinke y retirar el material sobrante a causa de la enfermedad.

Tras la operación, se recomienda a los pacientes un reposo vocal absoluto de 5 días tras el cual se pueden apreciar mejoras notables en la calidad de la voz.

1.4. Objetivos

El objetivo principal de este Trabajo de Fin de Grado es el de desarrollar un sistema que detecte a través de la voz si una persona padece o no la disfonía conocida como Edema de Reinke. Para llegar a dicho objetivo final, se han marcado objetivos intermedios de menor magnitud con el fin de conseguir una organización más eficiente:

- Desarrollo de una librería que reúna funciones para calcular los parámetros más utilizados para el estudio de disfonías.
- Extracción de parámetros de las X muestras de voz obtenidas a partir de la base de datos pública Saarbruecken Voice Database[SVD] (University, s.f.) y de las muestras del proyecto THALENTO, y su correcto etiquetado y agrupación según la vocal pronunciada en tres corpus distintos. Uno para la ‘a’ otro para la ‘i’ y otro para la ‘u’.
- Estudio estadístico y generación de gráficas sobre varios parámetros separados por sexo y vocal pronunciada.
- Estudio y análisis de los distintos algoritmos de aprendizaje automático que existen en la actualidad integrables en las tecnologías usadas durante el desarrollo.
- Diseño de dos clasificadores distintos utilizando el algoritmo de Gaussian Model Mixture.
- Diseño de una Red Neuronal Multicapa.
- División de cada Dataframe en conjuntos de Entrenamiento, Validación y Test.
- Entrenamiento de cada sistema con cada corpus.
- Ajuste de hiperparámetros y evaluación de resultados.

- Comparación de los tres tipos de clasificadores.

1.5. Herramientas

Para el desarrollo completo del trabajo he utilizado principalmente el IDE de Python, PyCharm Community, que permite el uso de entornos virtuales, tratamiento de errores mediante debugger e instalación sencilla de librerías.

También me he servido del software Praat, mundialmente conocido por la comunidad científica de las tecnologías del habla, y de su librería para Python, "Parselmouth".

2. Bases de datos

Para la realización de este estudio se han utilizado principalmente dos bases de datos:

2.1. Saarbrüecken Voice Database (SVD)

La SVD es una base de datos abierta creada por la universidad alemana de Saarlandes que contiene muestras de voz de más de 2000 personas divididas aproximadamente igual entre hombres y mujeres con más de 70 tipos de disfonías distintas. La interfaz web de esta base de datos puedes seleccionar tanto el rango de edad de los pacientes, el tipo de disfonía que padece, el sexo o incluso la fecha de grabación.

Por cada sujeto se encuentran los siguientes archivos:

- Grabaciones de los fonemas /a/, /i/, y /u/ cada una grabada de 4 formas distintas, una con voz neutra, una aguda, una grave y otra que varía entre agudo-grave-agudo.
- Los electroglotograma (EGG) de cada una de las 12 muestras mencionadas en el punto anterior.
- Una frase diciendo “Guten Morgen, wie geht es Ihnen?” que significa “Buenos días, ¿cómo estás?” y su correspondiente EGG.
- Notas sobre el diagnóstico del paciente.

Las muestras de voz vienen en formato “.wav” muestreado a 50kHz y 16bits por lo que son audios de buena calidad, incluso demasiada calidad para algunos de los parámetros que se han calculado, por lo que para el cálculo de algunos coeficientes se ha realizado un remuestreado a 16kHz.

Para este estudio, centrado únicamente en el ER se ha seleccionado a todos los sujetos etiquetados en la base de datos como pacientes de ER, extrayendo de ellos las muestras de voz neutrales, agudas y graves de las tres vocales. De todas las muestras antes comentadas se han excluido los electroglotogramas, porque no entraban en el rango del proyecto, las muestras de tono cambiante causaban una gran cantidad de errores en algunas estimaciones; y la frase, debido a que los parámetros que se calculan están preparados para hacerlos únicamente sobre vocales sostenidas.

En total, en la SVD hay 61 sujetos femeninos y 7 sujetos varones con disfonías diagnosticadas con edema de Reinke, así que para tener un set de datos equilibrado y no sobreajustar algoritmo de detección, se ha seleccionado 89 sujetos femeninos sanos y 43 sujetos masculinos sanos.

2.2. THALENTO

THALENTO (Tecnologias del HAbla y el Lenguaje para la EvaluaciónN de Trastornos de la cOmunicación) es un proyecto multidisciplinar que aúna los campos de la ingeniería, la medicina y la psico-educativa y logopédica con el objetivo de crear una gran base de datos abierta para la investigación sobre tecnologías del habla, que contenga tanto muestras de distintas disfonías como trastornos de comunicación.

Este proyecto comenzó en noviembre de 2019 impulsado por el tutor del trabajo Eduardo Lleida con la colaboración del grupo Otorrinolaringología del Hospital Clínico Universitario Lozano Bielsa dirigido por el Dr. Héctor Valles Varela, dicho proyecto se ha visto interrumpido a causa del COVID-19.

El estudio comenzó por la enfermedad del ER, así que a todos los sujetos diagnosticados con ER se les pedía su consentimiento para participar en el proyecto anónimamente obteniendo distintas grabaciones de su voz. En ese periodo de 4 meses se consiguieron grabar a 13 personas un tiempo antes de ser sometidos a la cirugía endolaríngea, de las cuales 12 son mujeres y sujeto restante, hombre. Las grabaciones posteriores a la operación solo pudieron ser realizadas a 3 sujetos.

Las grabaciones que se realizan a cada uno de los pacientes son:

1. Pronunciar las vocales /a/, /e/, /i/, /o/, /u/ de forma sostenida y aislada todo el tiempo posible, tres veces.
2. Soplar sostenidamente todo el tiempo posible (tres veces).
3. Pronunciar la consonante /s/ todo el tiempo posible (tres veces).
4. Pronunciar la consonante /z/ todo el tiempo posible (tres veces).
   Pronunciar la consonante /g/ todo el tiempo posible (tres veces).
5. Reproducir una escala musical pronunciando la /a/, /i/ y /u/ (tres veces).
6. Para cada vocal /a/, /i/ y /u/ pronunciarlas variando el tono (tres veces).
7. Lectura de un texto de unas 200 palabras
8. Conversación a partir de una lámina con preguntas abiertas.
9. Diálogo de 3 a 5 minutos con una de e las personas responsables de la grabación.

Por cada sujeto que accede a participar en el proyecto se obtienen los siguientes datos:

- Grabaciones de voz anteriormente comentadas.
- VHI 30 rellenado por el paciente durante el diagnóstico clínico.
- Encuesta ECV (Encuesta del comportamiento Vocal) que consta de 61 preguntas de múltiple respuesta sobre cómo la disfonía afecta a su vida diaria.

Para el desarrollo principal del trabajo solamente se han utilizado las muestras neutras, agudas y graves de las vocales sostenidas para así integrarlas con el resto de muestras de la SVD para crear una base de datos conjunta. Al no contar con las encuestas VHI ni las ECV de los sujetos de la SVD, no han sido utilizadas como entrada a la red neuronal. A pesar de ello, a continuación, analizaremos las encuestas y realizaremos un ligero estudio estadístico.

El conjunto de encuestas fue transcrito a un archivo "csv", el cual se ha utilizado para analizar los aspectos más comunes que comparten los pacientes de ER.

2.2.1. Encuestas ECV

Aquí mostraré los aspectos más comunes en pacientes de ER, al contar únicamente con 12 sujetos el estudio estadístico no es suficientemente preciso, pero hay ciertos aspectos que aportan información relevante. Las encuestas completas se pueden encontrar en el ANEXO I: Encuestas ECV.

2.2.1.1. Tabaco

Anteriormente hemos comentado que esta patología es habitual en fumadores, en la próxima gráfica vemos como el 75% de los pacientes fuma alrededor de un paquete o incluso un paquete y medio al día. (Figura 22)

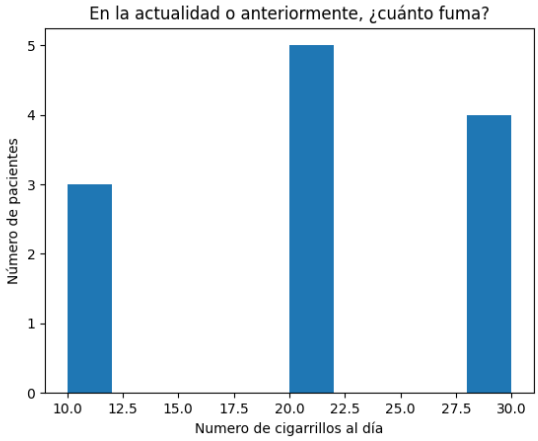

Figura 2. Histograma tabaco

2.2.1.2. Trastornos psico-emocionales

Numerosos estudios como por ejemplo en (Halawa et al., 2012) han demostrado como los antecedentes psicológicos como la ansiedad, estrés o depresión pueden afectar gravemente a la calidad de la voz. En la Figura 33 observamos que el 75% de los sujetos han contestado “sí” a tener trastornos emocionales.

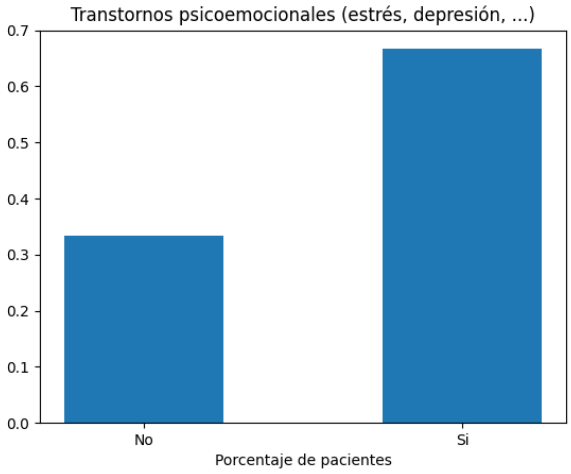

Figura 3. Histograma transtornos psicológicos

3.2.2. Encuestas VHI

La encuesta VHI se divide en tres partes: parte funcional, en la que se realizan preguntas sobre cómo te afecta la disfonía en las acciones comunes del día; parte física, donde se preguntan cuestiones relacionadas con la calidad de la voz; y la parte emocional donde las preguntas están relacionan con el impacto emocional que supone el no poder hablar con la comodidad que uno desea [ANEXO II: Encuestas VH].

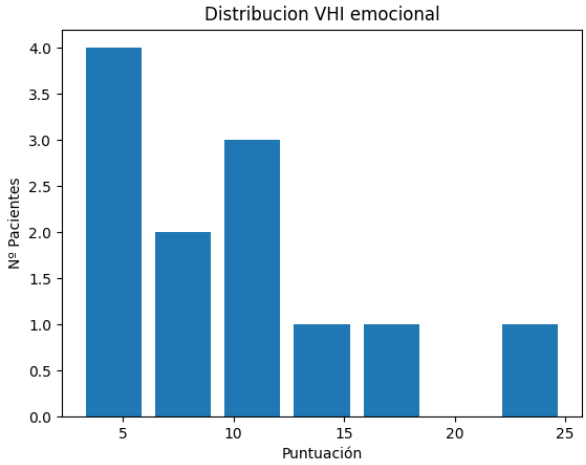

Figura 4. Histograma puntuación VHI emocional

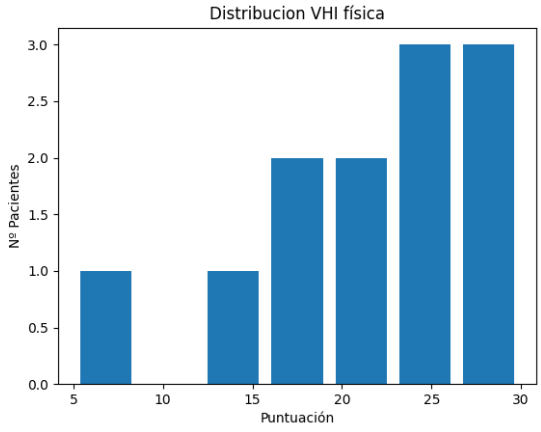

Figura 5. Histograma puntuación VHI física

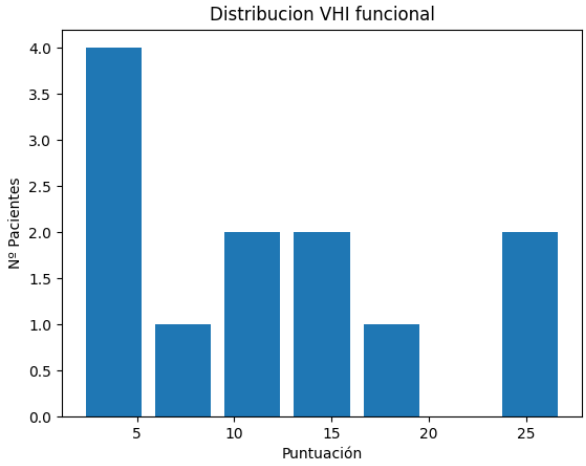

Figura 6. Histograma puntuación VHI funcional

Con estas gráficas (Figura 4 4, Figura 5 5, Figura 66), y teniendo en cuenta que el número de sujetos es muy bajo, se puede apreciar como no es tanto el impacto emocional o funcional sino el físico el que afecta gravemente a la vida del paciente.

3. Técnicas de procesado de señal

De este capítulo en adelante se comentarán los desarrollos teórico-prácticos realizados durante el transcurso del trabajo, así como los problemas que han surgido junto con sus correspondientes soluciones.

3.1. Extracción de parámetros

Tras haber explicado la procedencia de los datos que he utilizado, se explicaran a continuación los parámetros que se han extraído de cada una de las muestras de voz, las técnicas de procesado utilizadas y los algoritmos programados para obtener dichos valores. Para explicarlo dividiremos los parámetros en distintos grupos.

3.1.1. Frecuencia Fundamental

La frecuencia fundamental o frecuencia de pitch es la frecuencia a la que vibran las cuerdas vocales de forma estacionaria en cada persona, y que, junto a sus múltiplos, también denominados harmónicos, forman la composición completa de la voz humana.

La estimación de este parámetro ha sido un reto desde el comienzo de la telefonía, y gracias a su correcta estimación, es posible codificar la voz de manera muy eficiente. Por ello, a lo largo del tiempo se han desarrollado distintos algoritmos para el cálculo de este parámetro, a continuación, comentaremos las tres formas que se han utilizado para la extracción de este parámetro.

- Método de Autocorrelación: Consta de la división de la señal de voz en fragmentos de unos 40 ms y realizar la operación de autocorrelación sobre todos ellos para finalmente buscar la posición en la que se encuentra el máximo, dicha posición es la supuesta distancia en muestras entre dos periodos glotales. Este método es poco preciso ya que en multitud de ocasiones las señales de voz tienen más energía en el segundo o tercer harmónico que en el fundamental y finalmente no se ha utilizado en el desarrollo final.

- Método de Cepstrum: Consiste en realizar la operación de Cepstrum [1]a una señal dividida y enventanada en fragmentos de 40 ms, eliminar las muestras alrededor de 0 y buscamos la posición donde se encuentre el máximo. Esta posición en el eje de la quefrency [2]
- Software Praat: Este software mundialmente utilizado para el procesado de señal de voz tiene librerías para Python, las cuales he utilizado para extraer la frecuencia fundamental. El algoritmo que se utiliza es un algoritmo avanzado que utiliza técnicas de tracking junto al método de autocorrelación.

Tanto el método de Cepstrum como el que utiliza Praat son considerablemente precisos y en la mayoría de muestras de voz los dos algoritmos devuelven un resultado muy similar, pero en algunas circunstancias el algoritmo de Praat falla y devuelve un 0 o el de Cepstrum falla y devuelve la mitad de lo correcto por lo que para hacer la estimación más correcta posible aplico el máximo de los dos valores para el desarrollo siguiente, ya que la buena estimación de este valor va a ser determinante para el cálculo de algunos de los siguientes parámetros.

3.1.2. Jitter y Shimmer

Este grupo de parámetros fueron obtenido siguiendo el algoritmo desarrollado en (Teixeira & Gonçalves, 2016) , donde se trata de buscar los máximos valores y sus posiciones de cada periodo glotal (Figura 7). Dicho algoritmo está preparado para un gran número de tipos de voces, pero no para todas, ya que cada persona tiene una forma de onda propia, y en ocasiones pueden ser impredecibles para este tipo de algoritmos de búsqueda de máximos. Sobre todo, en este caso, en la que se procesan numerosas voces patológicas que habitualmente se caracterizan por ser altamente ruidosas e inestables, por lo que dificulta la estimación de las formas de onda.

Por esta razón, una parte considerable del tiempo total del trabajo ha sido dedicado a ajustar los parámetros detallados en el artículo y a añadir funciones complementarias

[1] $Cepstrum = F^{-1}\{\log F\{f(t)\}\}$
[2] Quefrency = Eje de abscisas de cepstrum

para que el algoritmo se ajustara a la mayor cantidad de tipos de formas de onda y evitar así posibles errores.

Hay numerosas muestras de voz que son altamente inestables y es muy difícil programar funciones para predecir y ajustar los puntos en los que se encuentran los pulsos glotales. Además, las muestras dependen considerablemente de la grabación y si el sujeto no la realiza bien o las circunstancias no lo permiten, ya sea por una molestia de voz externa a la enfermedad o cortes en la muestra.

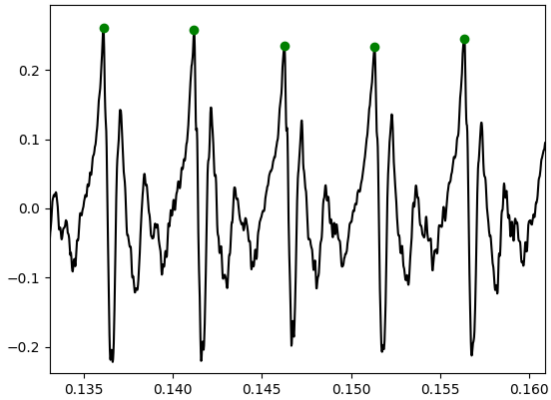

Figura 7. Búsqueda de máximos glotales

3.1.2.1. Jitter

El Jitter es la medida de las variaciones de periodo entre ciclo y ciclo, su origen procede de la falta de control sobre la vibración de las cuerdas vocales, por ejemplo, una de las enfermedades que más afecta a las cuerdas vocales es el Edema de Reinke. Para medirlo es necesario eliminar los tramos iniciales y finales ya que tienden a ser bastante inestables. Las perturbaciones de Jitter se calculan a través de 4 parámetros relacionados.

- Jitter(absoluto)-jitta: Es la diferencia media absoluta entre periodos consecutivos, expresado en μs.

$$jitta = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \qquad [1]$$

- Jitter(local)-jitt: Es la diferencia media absoluta dividida para el periodo medio, expresado en porcentaje.

$$jitt = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{l=1}^{N} T_i} \; x\; 100 \qquad [2]$$

- Jitter(RAP): Relative Average Perturbation es la diferencia media absoluta entre un periodo y sus dos vecinos, dividido por el periodo medio, expresado en porcentaje.

$$rap = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|\frac{1}{3}\sum_{n=i-1}^{i+1} T_n\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \qquad [3]$$

- Jitter(ppq5): five-point Period Perturbation Quotient (Cociente de perturbación de cinco puntos) es un parámetro prácticamente igual que el anterior, pero en vez de sus dos vecinos más próximos serán sus 4 vecinos próximos.

$$ppq5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|\frac{1}{5}\sum_{n=i-2}^{i+2} T_n\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \qquad [4]$$

3.1.2.2. Shimmer

Los parámetros de Shimmer están relacionados con la variación de amplitudes entre periodos consecutivos. Hay 4 parámetros de Shimmer:

- Shimer(dB) - ShdB: Es el logaritmo en base 10 del ratio medio absoluto entre amplitudes de periodos consecutivos. Medido en dB.

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}|20 * \log(A_{i+1}/A_i)| \qquad [5]$$

- Shimmer(local) – Shim: Es la diferencia media absoluta entre amplitudes de periodos consecutivos. Medido en porcentaje.

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{I=1}^{N} A_i}\ x\ 100 \qquad [6]$$

- Shimmer(APQ3): Cociente de perturbación de amplitud de tres puntos, es la diferencia media absoluta entre amplitudes entre un periodo y sus dos periodos adyacentes.

$$apq3 = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|\frac{1}{3}\sum_{n=i-1}^{i+1} A_n\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \qquad [7]$$

- Shimmer(APQ5): Cociente de perturbación de amplitud de cinco puntos, es la diferencia media absoluta entre amplitudes entre un periodo y sus cuatro periodos adyacentes más cercanos. Es decir, sus dos vecinos anteriores y sus dos vecinos posteriores.

$$apq5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|\frac{1}{5}\sum_{n=i-2}^{i+2} A_n\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \qquad [8]$$

3.1.3. Parámetros de ruido

Los parámetros relacionados con el ruido contienen una gran cantidad de información a la hora de la detección del edema de Reinke, ya que este ruido se suele deber a problemas en las cuerdas vocales, que son los que dan esa sensación de ronquera que caracteriza esta patología.

Son 6 los parámetros relacionados con el ruido que se van a explicar a continuación: Relación Ruido a Harmónicos (NHR), Relación Harmónicos a Ruido (HNR), Energía Normalizada de Ruido (NNE), Índice de Turbulencia de la Voz (VTI), Índice de Fonación (SPI) y ratio de la Excitación Glotal a Ruido (GNE).

- NHR: Es el ratio entre la energía que se encuentra entre las frecuencias de 1500 y 4500 Hz (frecuencias altas de ruido) y la energía de los harmónicos en un rango de 70 a 4500 Hz.
- VTI: Parámetro similar al anterior que trata de obtener las turbulencias de la señal en altas frecuencias. Se obtiene a través del cociente entre la energía de las componentes harmónicas que van desde los 2800 y los 5800 Hz, y las que van desde los 70 a 4500.
- SPI: Cociente entre la energía del harmónico fundamental y el harmónico de mayor frecuencia.
- HNR: Introducido por primera vez por Yamoto et al. (al, 1982) es seguramente el parámetro de ruido más estudiado en el ámbito del procesamiento de voz, ya que fue de los primeros en ser utilizados como diagnóstico de ronquera junto al Jitter y Shimmer. Se trata del ratio entre la energía total del ruido en la señal y la energía total de las componentes harmónicas. Su marco teórico se basa en que la señal de voz se puede explicar como la suma de una señal puramente periódica y un ruido aditivo de media 0. Así que siguiendo el algoritmo de la Correlación Normalizada (Boersma et al., 1993) podemos calcular el HNR a través de la siguiente ecuación.

$$HNR(dB) = 10 * \log_{10} \frac{r'(\tau_{max})}{1 - r'(\tau_{max})} \qquad [9]$$

Siendo $r'(\tau_{max})$ la autocorrelación normalizada en el punto T0.

- NNE: Desarrollado y estudiado por primera vez en el artículo de Kasuya et al. (1986) Se propone como una medida efectiva para la detección de patologías laríngeas, cáncer glotal, parálisis nerviosa recurrente y nódulos en las cuerdas vocales.

  El objetivo de este parámetro es bastante similar al del HNR, que es separar la señal entre señal periódica y señal de ruido, pero difieren en la forma de estimar la señal de ruido. Kasuya et al. (1986) dice que generalmente la señal de voz tiende a cambiar suavemente en amplitud y pitch a lo largo de la fonación sostenida, por lo que el método de Yumoto et al. (1982) podría detectar estos cambios incorrectamente como componentes de ruido. Este

método es teóricamente más robusto, frente a estos cambios de amplitud y pitch ya que el análisis es llevado a cabo utilizando ventanas de 7 periodos de pitch de longitud.

La fórmula utilizada para el cálculo del NNE y que explico a continuación es la siguiente.

$$NNE(db) = 10 * \log\left[\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{m=1}^{L}\left|\widehat{W}m(k)\right|^2 * \left(\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{m=1}^{L}|Xm(k)|^2\right)^{-1}\right] \qquad [10]$$

Donde N_L y N_H son las muestras correspondientes a la frecuencia más baja y a la más alta respectivamente de la banda frecuencial donde se quiere evaluar el ruido.

$\widehat{W}m$ es la estimación del ruido, cuyo cálculo depende de del tipo de banda frecuencial en la que se encuentre. Hay dos tipos, picos harmónicos y valles. En los picos la señal de ruido es calcula de la siguiente manera:

$$\left|\widehat{W}m(k)\right| = \frac{1}{2}\left\{\sum_{r\epsilon D_i}\frac{|Xm(r)|^2}{(N_{i-1})} + \sum_{r\epsilon D_{i+1}}\frac{|Xm(r)|^2}{(N_{i+1})}\right\}, k\epsilon P_{i-1}, P_{i+1} \qquad [11]$$

Y en las zonas de los valles:

$$\left|\widehat{W}m(k)\right| = |Xm(k)| \qquad [12]$$

- GNE: Introducido por primera vez por Michaelis et al (1996) como un nuevo parámetro para describir voces patológicas. Se basa en el coeficiente de correlación entre las envolventes de Hilbert de distintas bandas frecuenciales. En (D.Michaelis et al., 1996) demuestra como al contrario que el HNR y el NNE es casi totalmente independiente del error de modulación en frecuencia (Jitter) y de amplitud (Shimmer). El algoritmo utilizado para calcular este factor consta de los siguientes pasos:

1. Remuestrear a 10kHz.
2. Realizar un filtrado inverso mediante LPC (Linear Prediction Coefficients.
3. Calcular las envolventes de Hilbert de diferentes bandas frecuenciales con ancho de banda fijo y diferente frecuencia central.
4. Realizar la correlación cruzada normalizada entre cada posible par bandas frecuenciales diferentes y guardar el valor máximo de cada correlación.
5. Elegir el máximo valor de los máximos anteriormente encontrados.

3.1.4. MFCC: Mel Frequency Cepstrum Coefficients.

Los parámetros MFCC son unos coeficientes provenientes del dominio cepstral cuya característica principal es la gran cantidad de información sobre la voz que contienen. Son utilizados generalmente en sistemas de detección de voz y reconocimiento del interlocutor. Estos parámetros se distinguen de otros parámetros cepstrales por el tipo de banco de filtros que son aplicados, dichos filtros hacen uso de una nueva escala frecuencial no lineal denominada MEL que trata de imitar el comportamiento psicoacústico asignando mayor relevancia a las bajas frecuencias.

La obtención de estos coeficientes se lleva a cabo a partir de los siguientes pasos:

1. Enventanado. Se suelen utilizar ventanas de Hamming o Hanning del orden de 20-40ms con un solapamiento de entre 10 y 20 ms, de esta manera hacemos que la señal sea cuasi-estacionaria.
2. DFT. Se aplica la DFT del tamaño mínimo para que no se pierda ninguna muestra.

3. Banco de filtros. La señal de voz se filtra en el dominio frecuencial con un banco de filtros triangulares de área unidad espaciados respecto a la escala. (Figura 8)

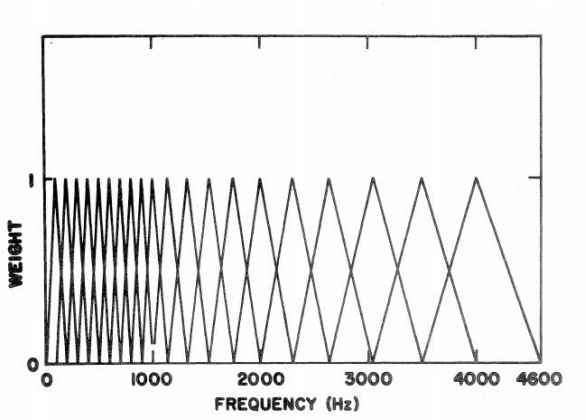

Figura 8. Banco de filtros mel utilizado por (Davis & Mermelstein, 1980)

4. A estos filtros triangulares se les aplica la transformación a la escala mel según la siguiente ecuación. (Figura 99)

$$mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad [13]$$

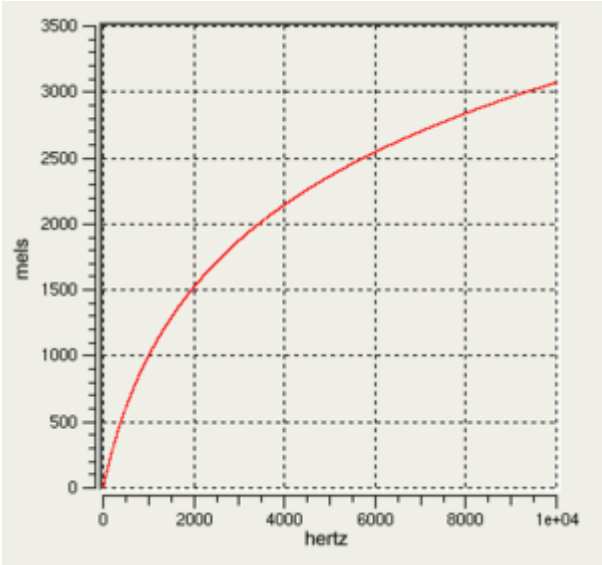

Figura 9. Escala de frecuencias mel

Donde f es la frecuencia en escala lineal.

5. Calcular el logaritmo de la energía correspondiente de cada uno de los filtros
6. Aplicación de la DCT para pasar al dominio Cepstral para finalmente convertirlos en coeficientes cepstrales.

Estos parámetros no han sido programados manualmente, sino que se ha utilizado la librería de código abierto Python_speech_features (Lyons, 2017) donde se puede ver cada uno de los pasos que se siguen para completar el cálculo de estos parámetros.

3.1.5. Cepstral Peak Prominence

El CPP o Prominencia del Pico Cepstral en español ha sido calificado como *la medida más prometedora y robusta para la severidad de la disfonía* (Y. Maryn et al, 2009). Fue introducido por primera vez por Hillenbrand et al. (1994) y ha sido un parámetro estudiado ampliamente por la comunidad médica como un parámetro muy relevante en la detección de ciertas disfonías. Merk et al. propuso la variabilidad del CPP como prueba para detectar desordenes vocales neurogénicos (1999), Rosa et al. Dijo que la combinación del CPP junto a otros parámetros acústicos es muy relevante en la detección de enfermedades laríngeas.

Su marco teórico se sitúa en el dominio cepstral, el cual permite descomponer la señal de voz en los distintos ecos que la forman, ya que la operación del logaritmo permite convertir el efecto multiplicativo de la señal moduladora proveniente del tracto vocal en un efecto aditivo (Godino et al., 2014).

Asumiendo la distinción entre tracto vocal de respuesta impulsional de longitud finita y señal glotal periódica, podemos diferenciar dos partes en el Cepstrum: la parte donde las quefrencies [3] son mayores que la quefrency fundamental ($q > q_0 = 1/f_0$) donde encontramos los rahmonicos[4] encontrados en múltiplos de q_0 y en las que es menor ($q < q_0$) donde podemos encontrar la transformación de la envolvente de las amplitudes de los harmónicos. En esta última parte es importante eliminar los primeros 2ms ya que allí es donde se sitúa la energía de la señal y es donde se nos pueden originar errores en el cálculo.

[3] Quefrency: Eje temporal del dominio cepstral, al igual que cepstral es un anagrama de "spectral" quefrency es un anagrama de frequency.
[4] Rahmonicos: Harmónicos en el dominio Cepstral

El procedimiento para calcular el CPP es el siguiente:

1. Cálculo del Cepstrum real utilizando una ventada de haming de 40ms con un solapamiento de 20 ms.
2. Suprimir los primeros 2ms de la señal.
3. Calcular la recta de regresión lineal del logaritmo del Cepstrum obtenido anteriormente, para obtener dicha recta he utilizado el método de mínimos cuadrados. (Figura 10)

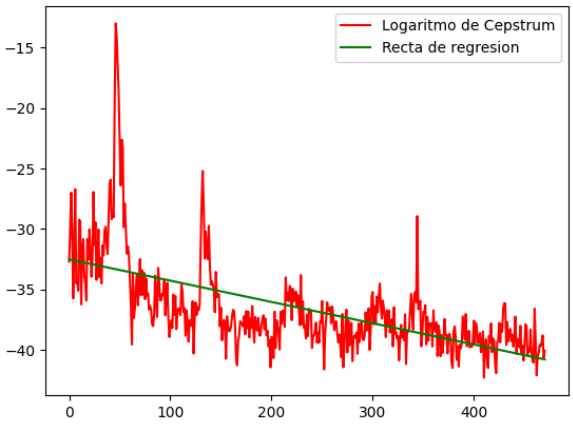

Figura 10 Logaritmo cepstral frente a su recta de regresión

4. Restarle la recta calculada en el paso anterior al logaritmo del Cepstrum.
5. Obtener el punto máximo de la señal calculada en 4. (Figura **¡Error! No se encuentra el origen de la referencia.**)

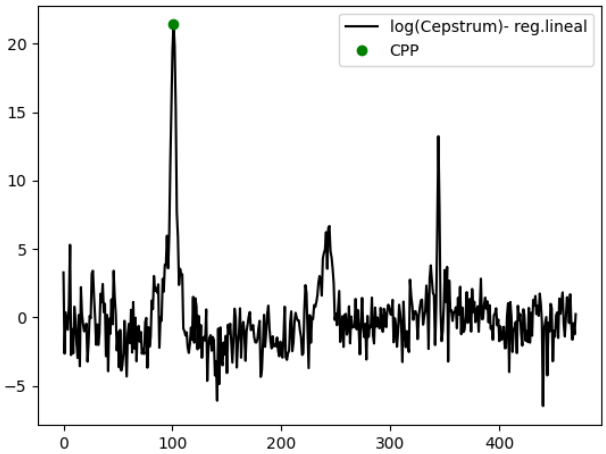

Figura 11. Obtención final del CPP

3.2. Estudio de clasificadores

En este apartado hablare de los dos algoritmos de clasificación que he utilizado para este trabajo: Gaussian Mixture Models y Redes Neuronales Multicapa.

El último tipo que se ha comentado, pertenece al grupo de algoritmos de redes neuronales, como su propio nombre indica, así que antes de plantear su desarrollo, se presentará una breve introducción explicando qué son las redes neuronales y cómo funcionan en general.

3.2.1. Gaussian Mixture Model

Un GMM, o Modelo de Mezcla de Gaussianas en español, es un modelo probabilístico que asume que los datos son generados a partir de una mezcla un número finito de distribuciones gaussianas con parámetros desconocidos. En el caso más sencillo, los GMM pueden ser utilizados para encontrar clústeres de la misma manera que hace el algoritmo de k-medias, el cual se basa en agrupar los sets de datos en clústeres circulares. Pero lo que nos ofrecen las GMM que no tiene el algoritmo de k-medias es que contienen un modelado probabilístico, el cual nos permite saber cuál es la probabilidad de cada punto a pertenecer a cada una de las distribuciones gaussianas.

En el campo del aprendizaje automático el algoritmo GMM forma parte de los algoritmos de aprendizaje no supervisado, los cuales se caracterizan por basarse en el agrupamiento de los datos en clústeres según su similitud, para así conocer mejor la estructura interna de los datos.

En este caso en concreto, nos interesa utilizar el GMM para dividir los datos en dos clústeres, y comprobar si las dos divisiones que haga el algoritmo coinciden con las etiquetas de los datos.

El GMM es uno de los más sencillos computacionalmente y a nivel de programación, además suele obtener buenos resultados con grupos de datos pequeños.

3.2.2. ¿Que son las redes neuronales?

Como todos sabemos, el cerebro es un gran procesador de información, capaz de procesar al mismo tiempo una gran cantidad de información a través de los sentidos, compararla con sensaciones similares anteriores y poder reaccionar ante ellas. Por ello, la comunidad científica lleva años estudiando el cerebro para modelarlo de la mejor manera posible, llegando al punto de las redes neuronales.

Las redes neuronales son una forma de modelar las neuronas y las conexiones entre ellas. Tratan de aprender y asociar patrones a partir de un determinado conjunto de entrados de entrada y de salida, por lo tanto, pertenecen al grupo de algoritmos de aprendizaje supervisado.

La unidad más básica dentro de las redes neuronales es la neurona artificial (NA), que modela a la neurona biológica, la cual se activa mediante una función de activación dependiendo de las entradas que recibe. La red más simple que se puede implementar es la del perceptrón simple (Figura 122).

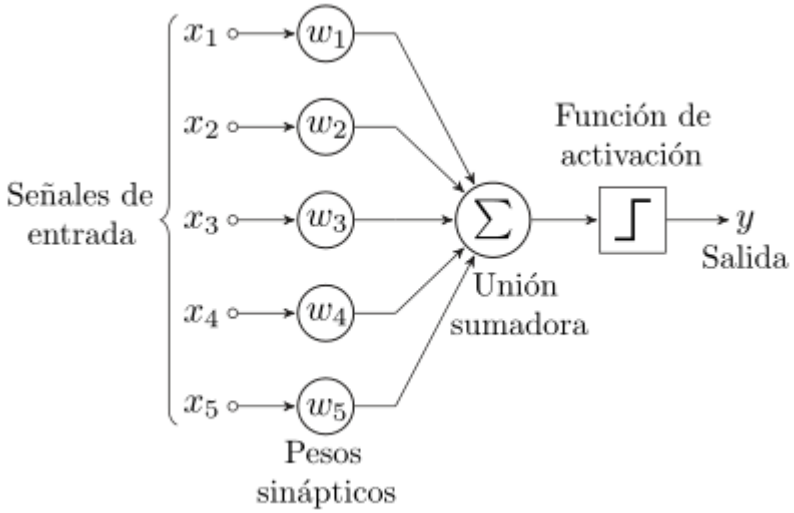

Figura 12. Ejemplo perceptrón Simple

En la Figura 122 podemos distinguir 3 operaciones principales: La multiplicación de cada entrada de datos por los pesos sinápticos, los cuales se actualizan conforme la red va aprendiendo, dando más importancia a las entradas con mayor cantidad de información. A continuación, la multiplicación por cada peso va a una unión sumadora. Finalmente, la sumarización de todas las entradas pasa por una función de activación, que transforman una entrada con valores en un rango específico en valores en rangos de (0,1) o (-1,1) dependiendo de la función.

3.2.2.1. Funciones de activación

Las funciones de activación son funciones no lineales que definen la salida de la neurona, generalmente tienen una derivada sencilla para reducir el coste computacional. A continuación, explicaré algunas de las más importantes.

- Rectified Linear Unit(ReLu): Es una de las más utilizadas en el campo del Deep Learning. La expresión de la función se puede escribir como $f(x) = \max(0, x)$, o sea, que da el valor de 0 a todas las entradas negativas y mantiene el de las positivas. (Figura 133)

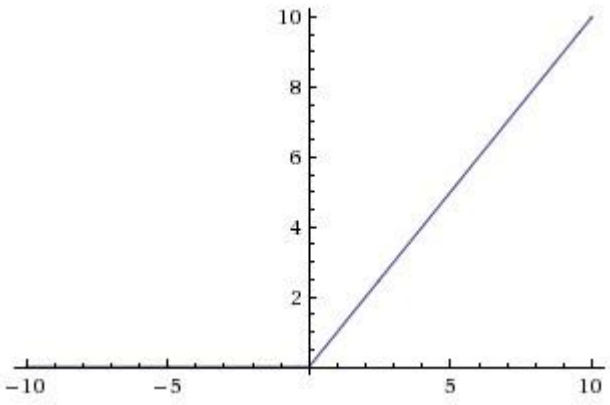

Figura 13. Función ReLu

- Sigmoide. La función sigmoide transforma los valores introducidos a una escala (0,1) donde los valores más bajos tienden asintóticamente a 0 y los valores más altos a 1. Es utilizada comúnmente en la última capa, donde se realiza la clasificación. La función se define como $f(x) = \frac{1}{1-e^x}$ y tiene la forma siguiente (Figura 144).

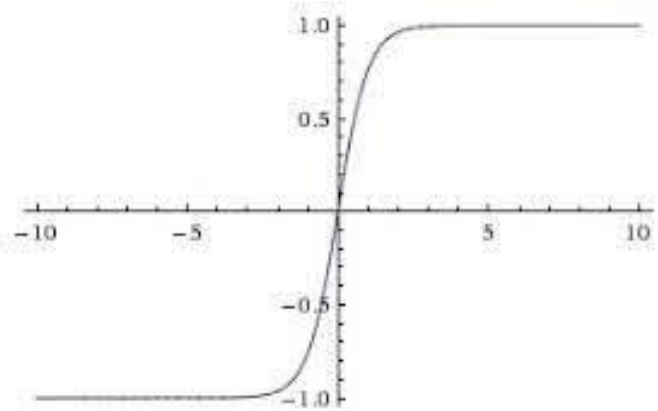

Figura 14. Función sigmoide

- Tangente Hiperbólica: Estas función transforma los valores introducidos a una escala(-1,1), la forma es muy similar a la función sigmoide pero centrada en 0. La función tanh(x) esta definida como $f(x) = \frac{2}{1+e^{-2x}} - 1$ y su aspecto es el siguiente (Figura 155).

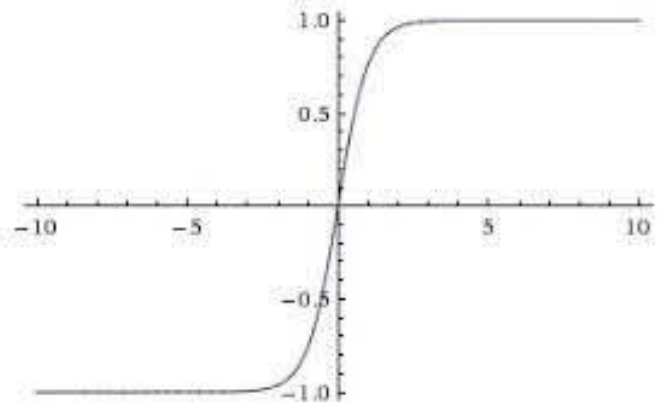

Figura 15. Función Tanh

- Softmax: Esta función transforma las salidas a una representación probabilística de tal manera que la suma de todas las probabilidades de las salidas es 1.

3.2.3. Redes Neuronales Multicapa

Las redes neuronales multicapa son un tipo de redes neuronales cuya estructura se caracteriza por estar formada por una secuencia de capas de NAs, en todas las redes multicapa hay como mínimo tres capas: Una capa inicial que tendrá tantas NA como entradas tenga el sistema; una capa oculta intermedia, y una capa final que tendrá tantas NA como salidas tenga el sistema (Figura 16). Utilizando el caso de este trabajo, en el que hay 30 parámetros de entrada, habría 30 NA en la capa inicial correspondientes a las 30 entradas de parámetros que tenemos el sistema, y dos neuronas en la capa de salida que representan las dos posibles salidas. Todas las NA de cada capa están conectadas completamente con cada neurona de las capas adyacentes.

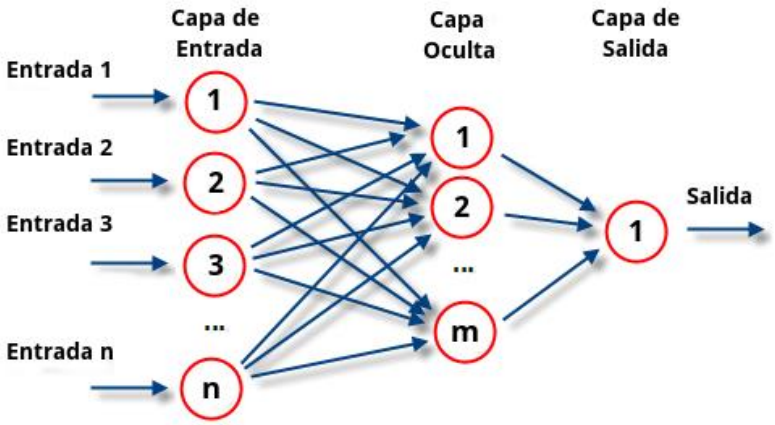

Figura 16. Esquema de red Neuronal Multicapa

Se pueden introducir tantas capas ocultas como se deseen, no por introducir más capas obtendremos mejores resultados, sino que debemos encontrar a base de pruebas la estructura que mejor se adapte a nuestro grupo de datos. Sí es cierto que cuanto más profunda sea la red y tenga más capas ocultas, será computacionalmente más costoso y se necesitarán una cantidad mayor de datos para que la red aprenda adecuadamente.

4. Diseño de sistemas de detección de ER

4.1. Creación de corpus de parámetros

Tras haber explicado detalladamente los distintos parámetros que utilizaremos en la detección, es el momento de organizarlos en el formato adecuado para poder utilizarlos posteriormente tanto para este trabajo en concreto como en otros estudios que se puedan realizar sobre este tema o de un tipo similar.

En este tipo de corpus, agruparemos a los parámetros que hemos calculado desde el punto 4.1.1 hasta el 4.1.5, es decir todos los parámetros que constan de un solo valor o una fila de valores concatenable con el resto de valores como es el caso de los parámetros MFCC, por lo tanto, contamos en total con 30 parámetros a los cuales les añadimos el sexo del sujeto y el nombre de la muestra. Por último, concatena la etiqueta de 0 o 1 según si es un sujeto sano o enfermo respectivamente.

Por lo tanto, nos encontramos con 33 valores a guardar por cada muestra de voz y se ira añadiendo cada fila de valores a una matriz de Nx33 siendo N el número de muestras a guardar en el archivo. Se han generado cuatro archivos: uno en el que aparecen los datos de todas las muestras de voz, y otros tres correspondientes a cada una de las tres vocales que son pronunciadas en las muestras. En uno sólo hay muestras en las que se pronuncia la /a/ en otra la /i/ y por último la /u/. De esta manera, podremos evaluar si alguna de las vocales contiene más información que las demás. El formato en el que se han guardado los corpus de parámetros ha sido en formato “xlsx” (Excel) a través de la librería pandas.

4.2. Diseño de clasificadores

Tras haber explicado los tipos de clasificadores que he tratado de utilizar en este trabajo, procedo a explicar de una manera más detallada como han sido programados cada uno de los tres clasificadores que figuran en el trabajo.

4.2.1. Clasificador GMM

Como he comentado anteriormente el apartado 3.2.1 el algoritmo GMM es un algoritmo de aprendizaje no supervisado que agrupa los datos en distintos clústeres según

sus características. El algoritmo GMM viene implementado en numerosas librerías de Python, para mi desarrollo he utilizado la librería “scikit-learn” que cuenta con la clase “GaussianMixture” en la que se ha basado este clasificador.

Como paso previo a todo sistema de ML, es necesario un acondicionamiento de los datos de entrada, que es la normalización de los datos, haciendo que cada una de las columnas de parámetros tengan media 0 y desviación estándar 1. Tras preparar los datos queda entrenar el algoritmo para que se adapte a los datos establecidos.

Utilizando el mismo algoritmo de GMM he utilizado dos enfoques a la hora de diseñar el clasificador para así poder compararlos y analizarlos.

La primera forma de diseñar el clasificador con GMM es utilizar el algoritmo como forma de agrupar en clústeres los datos en dos grupos, sanos y con ER. Para hacerlo de esta manera, utilizaremos un solo modelo con dos componentes, o sea, una mezcla de dos gaussianas. Este enfoque es el más sencillo de los dos ya que en realidad estamos utilizando el algoritmo de GMM simplemente para clusterizar prácticamente de la misma manera que lo hacen algoritmos más sencillos como el de K-Medias.

Antes de comenzar a entrenar el modelo es necesario definir el tipo de matriz de covarianza que modela cada función gaussiana. Las opciones son:

- Full: cada clúster tiene su propia matriz de covarianza
- Tied: todos los clústeres tienen la misma matriz de covarianza
- Diag: cada clúster tiene su propia matriz diagonal de covarianza
- Spherical: cada clúster tiene su propia varianza.

Para elegir el mejor tipo de covarianza posible se ha seguido el criterio BIC (Bayesian Information Criterion), un criterio utilizado en la selección de modelos basado en funciones de verosimilitud (Wikipedia, s.f.). A continuación, se muestra un gráfico comparando como cambia el BIC dependiendo del tipo de matriz de covarianza que se utiliza y el número de componentes. Para el entrenamiento de estos modelos se utilizado el grupo de sujetos sanos del corpus completo DF.

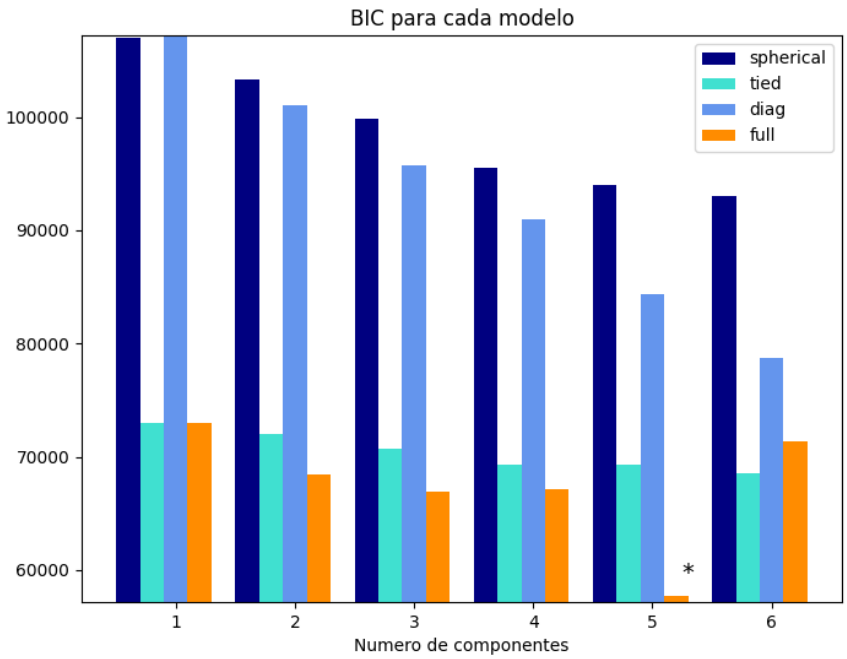

Figura 17. Comparación tipos de covarianza

Mirando la Figura 177, se podría decir de antemano que el modelo que mejor modela el grupo de sujetos sanos es un modelo de mezcla de 5 gaussianas con función de covarianza 'Full'.

Tras haber definido la instancia de 'MixtureModel' es el momento de entrenar el modelo mediante la función "fit_predict()" que además de entrenar el modelo con la secuencia de datos que se le introduce como parámetro devuelve el clúster donde ha situado a cada grupo de datos, así podemos comparar dicha salida con las etiquetas verdaderas. Para el entrenamiento se ha utilizado el 80% del grupo de datos para el entrenamiento y el 20% restante para test. Tras haber entrenado el sistema se le introducen los datos de test mediante la función 'predict()' y se compara la salida con las etiquetas reales para calcular la probabilidad de error.

El segundo enfoque que se ha desarrollado, y que es el enfoque habitual que se le suele dar al algoritmo de GMM, es el de crear varios modelos, tantos como etiquetas distintas hayan, formados por un número de gaussianas determinado, que modelen cada uno de los grupos existentes en el corpus. En este caso se crean dos modelos uno para modelar los sujetos sanos y otro para modelar los sujetos con ER. Tras modelar cada

grupo de datos se introducirá cada dato a cada uno de los modelos y se calculará la probabilidad de pertenecer a cada modelo, después se comparan las probabilidades y donde la probabilidad sea mayor se le asignara la etiqueta perteneciente a ese modelo.

Para desarrollar este clasificador, se ha comenzado definiendo los dos modelos, asignándoles al principio dos componentes a cada uno y tipo de covarianza 'Full'. Después, se entrena uno de los modelos solamente con los datos de los sujetos sanos. Así tendremos un conjunto de gaussianas que modelen lo mejor posible a los sujetos sanos. Tras entrenar el primer modelo, guardamos el vector de medias y el vector de pesos de ese modelo, e inicializamos el segundo modelo con el mismo vector de medias y el mismo vector de pesos. De esta manera, el modelo partirá del mismo punto que el modelo de sujetos sanos, y reentrenándolo con sujetos con Reinke captará las diferencias existentes y esas medias y pesos se moverán hacia las medias correspondientes de los sujetos con ER.

Es necesario aclarar que para el entrenamiento de cada modelo, se ha utilizado el 80% de su corpus correspondiente, y el otro 20% se ha utilizado para el test.

Tras tener el sistema completo desarrollado, mediante prueba y error, he seleccionado el número de componentes con el que los modelos obtienen la menor probabilidad de error posible.

4.2.2. Clasificador RN Multicapa

En este apartado de presenta de forma detallada la estructura de la red neuronal multicapa y como ha sido entrenada.

Para el entrenamiento de esta red neuronal hemos utilizado únicamente los corpus de parámetros.

Al igual que en el sistema anterior, antes de introducir los datos en la red es necesario normalizarlos para que tengan media 0 y desviación estándar 1, de esta manera la red neuronal trabaja en todas las neuronas en el mismo rango de valores y aprende de una forma más eficiente, tras la normalización se dividen en tres grupos, uno de entrenamiento, otro de test y otro de validación, en la proporción de 70%, 15% y 15% respectivamente.

Tras la normalización, es el momento de definir las capas de la red neuronal. Para comenzar, creamos la capa que sirve de interfaz de entrada de datos a la red neuronal, como se ha comentado en el apartado de 4.1, nuestro corpus tiene un total de 30

parámetros por cada muestra de voz por lo que esta capa inicial contará con 30 neuronas. Tras definir la capa de entrada es el momento de definir las capas ocultas. El número de capas y el número de neuronas por cada capa han seguido un proceso de prueba y error para encontrar la red que mejor se adapte a los datos.

La primera capa oculta consta de 32 neuronas que se activan mediante la función de activación “Tangente Hiperbólica” ya que tenemos como entrada parámetros con valores tantos positivos como negativos, por lo que colocar otra función como por ejemplo la ReLu nos haría perder mucha información.

Las dos siguientes capas utilizan la función de activación Relu, ya que funcionan muy bien en capas intermedias. La primera capa consta de 16 neuronas y la segunda de 8.

Por último, en la capa final contamos con una capa de una única neurona que se activa mediante la función ‘sigmoid’ y es la que determina la clasificación final.7

Para el entrenamiento se ha utilizado el optimizador Adam con un ratio de aprendizaje de 0.0005 fijado tras probar distintos valores.

5. Resultados

Este capítulo final se dividirá en tres partes:

Una primera parte demostraremos gráficamente mediante el espectrograma las diferencias espectrales entre el preoperatorio y el postoperatorio.

Una segunda parte en la que se analizaran estadísticamente los parámetros calculados y veremos las diferencias objetivas entre sujetos enfermos sanos y sujetos enfermos, separados por sexo e incluso por vocal pronunciada. Además, contando con los sujetos de THALENTO, veremos las significativas diferencias que hay entre las muestras antes de la operación y las muestras después de la operación.

Después, en la última parte analizaremos los resultados de los clasificadores y los compararemos entre ellos.

5.1. Espectrograma

El espectrograma es una representación tiempo-frecuencia de una señal basada en la transformada de Fourier localizada.

El cálculo se realiza mediante el enventanamiento de la señal en N ventanas de corta duración M y solapadas entre sí. Después, a cada ventana se le aplica una transformación frecuencial como la FFT. El último paso trata de unir todas las ventanas en forma de una matriz MxN donde se puede apreciar la evolución temporal de la distribución frecuencial de la energía en la señal de voz.

En el espectrograma se puede visualizar fácilmente como de ruidosa es una señal. Como ejemplo, a continuación, se muestra la comparación entre el espectrograma de uno de los sujetos de THALENTO antes de la cirugía, y el espectrograma del mismo sujeto después de la cirugía. Para el cálculo se ha realizado un remuestreo a 16kHz, aplicado una ventana de hanning de 1024 puntos con un solapamiento de 128 puntos.

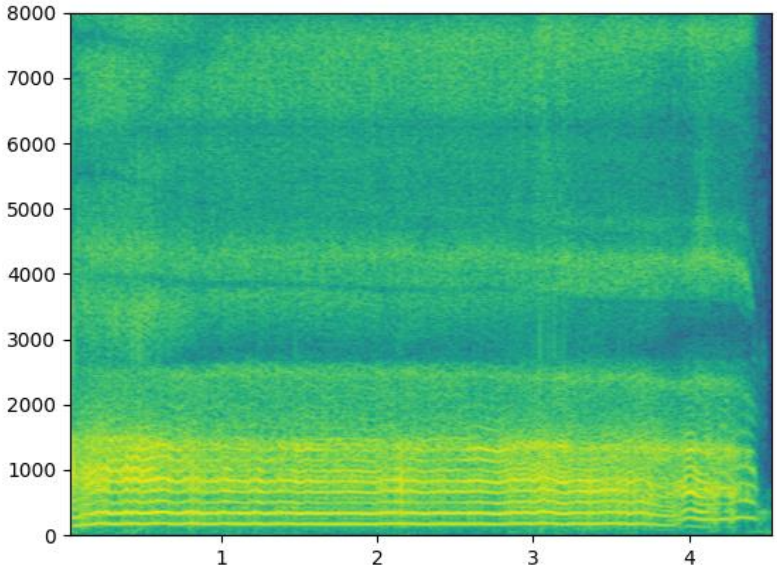

Figura 19. Espectrograma pre-operatorio

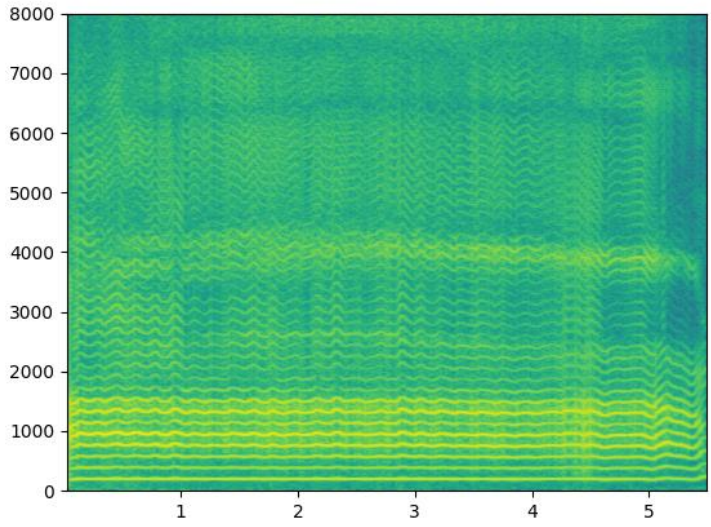

Figura 18. Espectrograma post-operatorio

En la Figura 198 podemos visualizar los harmónicos de baja frecuencia, pero de una forma considerablemente difuminada a causa del ruido. Además, fijándonos en las altas frecuencias casi no se distinguen los harmónicos, solo vemos una zona difuminada por el ruido, que es este caso tiene incluso más energía que las componentes periódicas.

En cambio, en la Figura 189 apreciamos unas marcadas componentes periódicas correspondientes a los harmónicos, tanto en baja frecuencia, donde se sitúan los harmónicos con más energía, como en alta frecuencia, donde las componentes de ruido suelen tener más peso que las anteriores.

5.2. Análisis estadístico

En el apartado 3.1 se ha comentado cada uno de los parámetros calculados en el desarrollo del proyecto. Ahora, analizaremos e interpretaremos los resultados obtenidos de dichos cálculos. Nos aprovecharemos de la función "describe()" de la clase "pandas.Dataframe" que nos proporciona los siguientes:

1. Count: número de filas analizadas.
2. Mean: media aritmética.
3. Std: desviación estándar.
4. Mínimo y máximo.
5. Percentiles del 25%, 50% y 75%.

Para comparar los datos se han dividido los corpus por sexo y por sano/patológico para ver las diferencias entre ellos. Para compararlos he utilizado las letras N y R para distinguir entre voz sana y enfermo respectivamente, y las letras M y F para distinguir entre masculino y femenino.

- **Frecuencia de pitch.**

F0 Total				
	MN	MR	FN	FR
count	387	72	827	656
mean	141,33	134,63	243,65	184,42
std	46,15	63,27	51,08	53,69
min	59,42	66,39	108,84	71,43
25%	108,94	107,38	213,33	143,33
50%	128,66	116,38	238,91	181,82
75%	161,62	145,45	266,67	213,33
max	367,29	405,65	466,98	421,05

Tabla 1: Estadísticas F0

En la Tabla 1 podemos verificar la afirmación de que la frecuencia fundamental es más baja en sujetos con edema de Reinke. Se puede destacar como la diferencia aumenta considerablemente en mujeres. Esto puede ser debido a que tienen una frecuencia fundamental media mayor y la ronquera que produce el ER provoca una bajada del tono mayor. La diferencia no es solo perceptible en la media sino también en todos los percentiles.

El dato marcado en rojo puede resultar confuso, ya que esa f0 no se alcanza en sujetos masculinos en ningún caso, además fijándonos en las tablas de estadísticas separadas por vocales (ANEXO III: Tablas de estadísticas por vocal) vemos que este valor solo se alcanza en la vocal "a", esto es debido a algunos de los errores que produce el algoritmo de estimación de frecuencia fundamental.

- **Jitter.**

Jitter Total				
	MN	MR	FN	FR
count	387	72	827	656
mean	1,24	2,37	1,75	3,24
std	1,25	2,62	1,26	2,58
min	0,18	0,42	0,25	0,30
25%	0,51	0,71	0,90	1,19
50%	0,85	1,49	1,36	2,40
75%	1,49	2,88	2,19	4,80
max	9,99	14,78	9,97	16,50

Tabla 2 Estadísticas Jitter.

Con los datos de la Tabla 2 podemos comprobar la relevancia del Jitter en el campo de la detección de disfonías. En (Teixeira et al) se fija como valor límite entre voces patológicas y sanas 1.04, clasificando como sanas a las voces con Jitter por debajo de ese valor y patológicas por encima. En éste caso, ese valor está por debajo de cualquier media de sujetos sanos por lo que no serviría. Tampoco es justo asignar un valor fijo para todos los tipos de voces ya que hombres y mujeres tienen por defecto características vocales distintas. Podemos observar diferencias notables en todos los percentiles, sobre todo en los casos de sujetos femeninos, también es cierto que al tener mucha más información de sujetos femeninos contamos con una información más representativa. Al igual que en el caso anterior, en los valores máximos vemos valores muy grandes

comparados con lo habitual, es posible que se deban a un verdadero Jitter muy extenso, pero es altamente probable que se deba a un error del algoritmo (Teixeira & Gonçalves). Aunque los parámetros de jitta, apq3 y apq5 son relevantes y son utilizados en los clasificadores, no se ha realizado un estudio individual ya que están directamente relacionados con el Jitter (%).

- **Shimmer**.

Shimmer Total				
	MN	MR	FN	FR
count	387	72	827	656
mean	3,53	12,46	5,48	11,23
std	3,33	43,85	3,98	17,57
min	0,50	0,49	0,66	0,29
25%	1,69	2,10	2,81	2,97
50%	2,57	3,79	4,47	6,07
75%	4,24	6,24	7,02	11,49
max	32,49	270,50	40,52	201,91

Tabla 3 Estadísticas Shimmer

De nuevo se pueden apreciar notables diferencias entre los pacientes de ER y los sujetos sanos, tanto en el sexo femenino como masculino. En (Teixeira et al) se fija el valor de Shimmer en 3.81 para distinguir a las voces entre patológicas y sanas. Fijándonos en el percentil del 50% (Tabla 3), ya que la media puede estar contaminada por los valores tan grandes que se producen por errores del algoritmo, este valor podría ser utilizado en el grupo masculino de la vocal ‘a’, o en el grupo femenino de la vocal (ANEXO III: Tablas de estadísticas por vocal. Como podemos comprobar, tanto en este parámetro como en los anteriores, hay diferencias considerables entre las muestras de la misma persona pronunciando vocales distintas. De ahí la idea de generar modelos distintos para cada vocal.

Tras haber analizado los parámetros de F0, Jitter y Shimmer, procederemos a analizar las estadísticas obtenidas de los parámetros de ruido obtenidos.

- **HNR**.

HNR Total				
	MN	MR	FN	FR
count	387	72	827	656
mean	25,81	21,19	26,94	16,75
std	5,25	8,75	5,07	14,41
min	-9,97	0,02	-7,77	-116,07
25%	23,15	15,44	24,24	12,59
50%	26,13	21,87	27,11	20,15
75%	28,98	27,43	30,25	25,32
max	37,92	36,85	41,88	40,34

Tabla 4 Estadísticas HNR

En (Teixeira et al) se fija el valor de 7 dB como límite entre voces patológicas y voces sanas, pero en la Tabla 4 observamos como la gran mayoría de los valores están por encima de este valor. Eliminando los datos de mínimos y máximos, se puede apreciar como hay más o menos una diferencia de 5 dB entre las voces sanas y las voces patológicas. A continuación, se muestra los histogramas de los valores de HNR de las muestras de mujeres ya que el conjunto de datos es más representativo que el de los hombres.

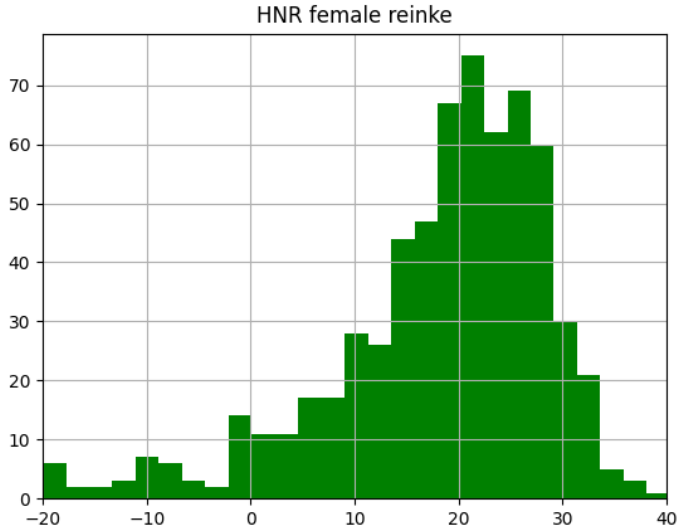

Figura 20. Histograma HNR FR

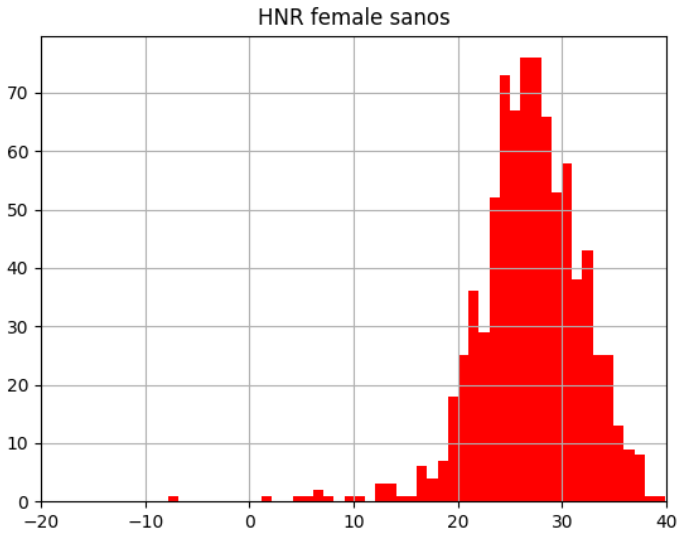

Figura 21. Histograma HNR FN

Con los histogramas presentados en la Figura 2020 y la Figura 211 podemos observar que el histograma de sanos tiene una gorma gaussiana mucho más marcada que la de sujetos con ER. Aun así, se puede apreciar la diferencia de medias entre las dos distribuciones.

- **NHR**.

NHR Total				
	MN	MR	FN	FR
count	387	72	827	656
mean	-33,35	-31,70	-34,75	-33,57
std	11,15	10,68	9,86	11,01
min	-85,65	-52,31	-88,82	-90,00
25%	-40,91	-39,37	-42,47	-40,93
50%	-31,94	-30,08	-33,19	-32,26
75%	-25,93	-23,26	-28,06	-26,57
max	-8,35	-11,47	-9,57	-6,45

Tabla 5 Estadísticas NHR

En este caso, fijándonos en la Tabla 5, no se pueden apreciar diferencias notables entre los sujetos con ER y los sujetos sanos. Sí que es cierto que se puede ver como en la mayoría de los percentiles los valores son ligeramente menores en los sujetos sanos que en los sujetos con ER, pero no es una diferencia suficientemente significativa como para destacarla.

- **GNE**

GNE Total				
	MN	MR	FN	FR
count	387	72	827	656
mean	0,83	0,76	0,75	0,70
std	0,15	0,15	0,16	0,17
min	0,28	0,29	0,24	0,22
25%	0,75	0,69	0,64	0,57
50%	0,88	0,77	0,79	0,71
75%	0,94	0,88	0,89	0,83
max	0,99	0,97	0,99	0,98

Tabla 6 Estadísticas GNE

El GNE es un parámetro cuyo valor siempre se encuentra entre 0 y 1, y como podemos observar en la Tabla 6, sí que se aprecian diferencias notables entre sujetos con ER y sujetos sanos en los percentiles del 25, 50 y 75%.

- **NNE**.

NNE Total				
	MN	MR	FN	FR
count	387	72	827	656
mean	-3,94	-3,21	-2,62	-2,36
std	2,94	2,27	3,77	2,53
min	-18,16	-9,36	-16,28	-12,66
25%	-5,46	-4,07	-4,69	-3,40
50%	-3,52	-2,83	-3,02	-2,20
75%	-2,15	-1,72	-0,72	-0,99
max	13,31	2,74	14,41	7,38

Tabla 7 Estadísticas NNE

En este parámetro, se debería apreciar una clara diferencia entre los sujetos con ER y los sujetos sanos, siendo más altos en los casos con ER, ya que, como se ha comentado anteriormente, las voces patológicas tienen más energía de ruido que las voces sanas. En cambio, fijándonos en los percentiles de la Tabla 7 no vemos ninguna diferencia notable entre los sujetos con ER y los sujetos sanos.

- **CPP**.

CPP Total				
	MN	MR	FN	FR
count	387	72	827	656
mean	19,61	16,30	20,04	14,68
std	3,43	3,93	3,16	4,85
min	9,95	8,31	7,98	2,55
25%	17,43	12,54	17,78	11,34
50%	19,84	16,84	20,04	15,07
75%	21,90	19,02	22,38	18,28
max	27,99	26,49	28,25	28,85

Tabla 8 Estadísticas CPP

Analizando las estadísticas, comprobamos que el CPP es el parámetro más estable en cuanto a errores de cálculo por culpa del algoritmo ya que en ninguno de los máximos ni mínimos observamos valores muy distanciados de la media. Además, a simple vista se puede ver que en todos los percentiles hay como mínimo 2,5dB de diferencia entre los DS de sujetos sanos y los de sujetos con ER.

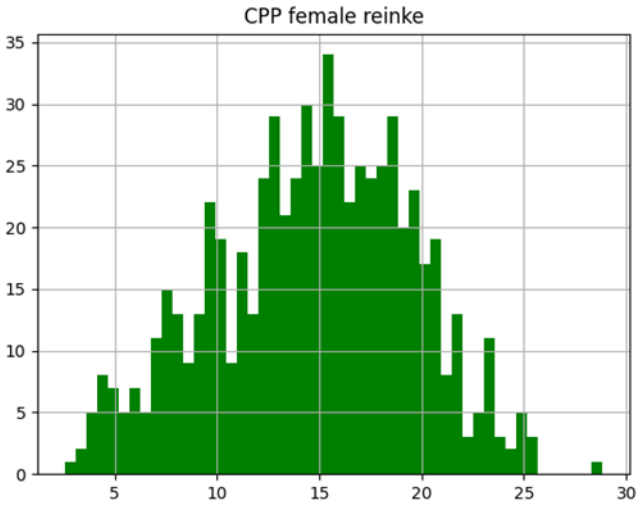

Figura 22. Distribución de CPP en sujetos femeninos con ER

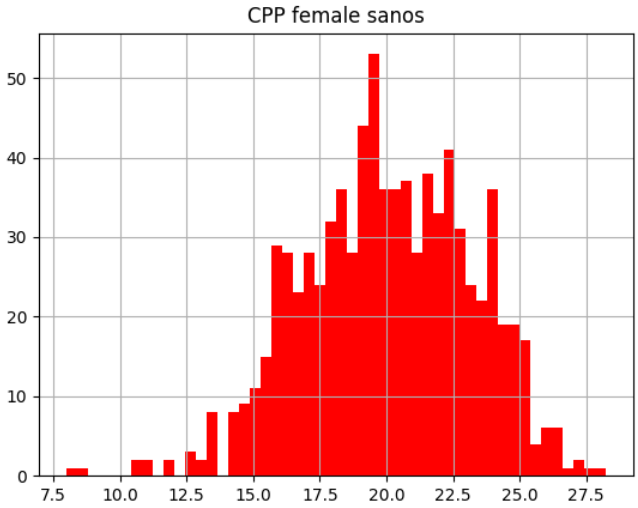

Figura 23. Distribución CPP de sujetos femeninos sanos

Visualizando los histogramas podemos ver como los datos siguen una distribución similar a una función gaussiana. Ambos histogramas tienen una forma parecida pero el primer histograma (Figura 222**¡Error! No se encuentra el origen de la referencia.**), está centrada en 15dB y la segunda (Figura 233) en 20dB. Realmente, las gaussianas están considerablemente solapadas, pero la diferencia en la media de unos 5 dB nos da una información muy valiosa.

5.3. Clasificadores

En esta parte final del trabajo se mostrarán las precisiones obtenidas por los clasificadores a la hora de detectar si el sujeto padece ER o no.

Para cada tipo de sistema, se han realizado modelos distintos para cada uno de los corpus. Además, para cada sistema y corpus se han realizado 10 pruebas distintas para poder obtener medidas más realistas sobre la detección, ya que es posible con un grupo de datos de entrenamiento específico se obtengan resultados que no se ajusten del todo a la realidad. Como el mezclado previo a la separación entre grupo de entrenamiento y grupo de test es aleatorio, se hacen estas 10 realizaciones para así tener más datos que estudiar.

Tras hacer las 10 realizaciones de entrenamiento de cada algoritmo y guardando el valor de precisión en el conjunto de test, se calcula el valor medio de esas 10 realizaciones, la desviación estándar, el valor mínimo y el valor máximo. Con estos valores, creamos una tabla de Excel donde las filas son los valores calculados nombrados anteriormente, y las columnas son los corpus de los cuales se han calculado esos valores y están nombradas como "DF_vocal". utilizando el acrónimo "DF" refiriéndose a Dataframe que se podría traducir como grupo de datos. La columna "DF" e refiere al grupo de datos que aúna las muestras de todas las vocales

A continuación se muestran los resultados del primer sistema desarrollado: el uso de GMM para agrupar los datos en clústeres según sus características.

GMM 1	DF_A	DF_I	DF_U	DF
media	73,32	68,55	74,72	69,36
std	3,76	8,79	4,32	6,81
min	61,54	50,77	57,69	53,21
max	80,77	79,23	83,85	76,61

Tabla 9. Resultados GMM, Sistema 1.

Con la Tabla 9 podemos confirmar algo que se ha comentado anteriormente, que los resultados dependen considerablemente del conjunto de datos con los que se entrena y con el que se entrena el modelo. Por ejemplo, vemos como en el grupo DF_I hay una diferencia de 30 puntos porcentuales entre el mínimo y el máximo, siendo este valor mínimo un 50% de precisión, lo que se traduce en el peor resultado posible en una clasificación con dos etiquetas.

Por suerte, este tipo de resultados son excepcionales ya que vemos que la media de la precisión en cada corpus va desde el 68,55% en el DF_I, hasta un 74,72% en el DF_U. Además, vemos que se pueden alcanzar valores superiores al 80% como en el caso del DF_A y el DF.

Estos resultados son considerablemente buenos contando con la simplicidad de este algoritmo. Ya que este sistema es el más sencillo que se ha desarrollado en este trabajo, estos resultados son los que se tratan de mejorar con los otros dos sistemas.

A continuación, se mostrarán los resultados obtenidos con el segundo sistema desarrollado con el algoritmo de GMM, en el que se crea un modelo de mezcla de

gaussianas, con un número determinado de funciones gaussianas, para cada grupo de datos. En este caso se han hecho dos modelos, uno para ER y otro para sanos.

Para ver qué número de componentes es el óptimo para estos modelos, se han hecho 9 sistemas distintos incrementando en uno cada vez el número de componentes que tienen los modelos y al igual que se ha hecho antes, se ha entrenado cada sistema 10 veces para poder compararlos de una manera equilibrada y así poder decidir cuál o cuáles son los valores óptimos. Para mostrar la información más relevante, se muestran únicamente los resultados del sistema diseñado con 4 gaussianas en cada modelo (Tabla 10), el resto de resultados se encuentran en el ANEXO IV: Tablas de resultados GMM.

GMM 4	DF_A	DF_I	DF_U	DF
media	88,87	91,13	90,31	90,92
std	5,10	1,49	3,35	1,09
min	78,97	88,21	85,13	87,84
max	97,44	92,82	95,38	91,78

Tabla 10. Resultados GMM, Sistema 2.

Cada tabla corresponde a cada sistema GMM desarrollado con dos modelos cada uno, cuyo número de componentes es el número que aparece en cada tabla.

Fijándonos en el corpus completo 'DF', no vemos diferencias entre los distintos sistemas, todas cuentan con valores cercanos al 90%, se podrían destacar los sistemas con 7 y 8 gaussianas cuyos valores medios de precisión son de 94.02% y de 93.52% de acierto.

En cambio, fijándonos en los corpus en los que se pronuncia la misma vocal, 'DF_A', 'DF_I' y 'DF_U', la precisión varía considerablemente cambiando el número de componentes con las que están construidos los modelos. Por ejemplo, podemos comparar el GMM modelado con 4 componentes y el modelado con 10, ambos modelos tienen una media de precisión en el corpus completo DF del 90% pero en estos mismos modelos fijándonos en los resultados obtenidos en los otros corpus, la diferencia aumenta considerablemente. En el GMM4 los modelos entrenados con los grupos específicos tienen una eficiencia media como mínimo del 88,87%, mientras que en el GMM10 el modelo con mayor precisión media es del 76,82%. Además, en el GMM4 con el grupo DF_A se ha llegado a obtener una precisión del 97,43% una precisión muy alta teniendo en cuenta la cantidad limitada de datos de entrenamiento con la que cuenta el sistema.

Por último, se mostrarán los resultados de la Red Neuronal Multicapa, la cual ha sido sometida el mismo procedimiento que los otros sistemas para comparar los resultados entre diferentes corpus y diferentes realizaciones.

MNN	DF_A	DF_I	DF_U	DF
media	84,43	84,12	80,00	84,40
std	2,59	3,07	2,45	1,82
min	80,41	78,35	75,26	80,41
max	88,66	89,69	85,57	86,94

Tabla 11. Resultados MNN.

A partir de la Tabla 11, podemos ver que las precisiones de la red neuronal diseñada con los estos datos se sitúan alrededor del 80% destacando algunos valores máximos como el valor máximo de las realizaciones con el grupo 'DF_I' que alcanza casi el 90 % de precisión. Comparados con los resultados del GMM mostrados anteriormente, podemos observar que son ligeramente peores. Esto puede ser debido a que el conjunto de datos no es los suficientemente grande para que aprenda las características de los datos adecuadamente. Habitualmente, para entrenar una red neuronal se utilizan cantidades ingentes de datos de tal manera que así la red pueda aprender el máximo número de casos reales posibles.

6. Conclusiones

En este estudio, se ha hecho un repaso de las características médicas y acústicas de las voces de personas diagnosticadas con Edema de Reinke. Se ha analizado algunos de los hábitos que comparten los pacientes de esta enfermedad tales como el hábito de fumar, la exposición continua de la voz en trabajos donde su uso sea prolongado e incluso los problemas psicológicos, extrayendo así información relevante para la prevención anticipada del Edema de Reinke.

Por otra parte, se ha desarrollado un conjunto de librerías donde se encuentran las funciones necesarias para calcular los principales parámetros acústicos en el campo del procesamiento de señal de voz, de tal manera que pueda ser utilizada en el futuro por el proyecto THALENTO, para realizar diferentes estudios sobre otras patologías de la voz.

Posteriormente, con las estadísticas realizadas sobre los parámetros, se han podido demostrar las diferencias cuantificables entre voces patológicas y sanas.

Además, haciendo uso de las escasas muestras de la base de datos de THALENTO hemos podido ver de manera objetiva las diferencias entre la voz en el preoperatorio y el postoperatorio. Gracias a esto, podemos llegar a la conclusión antes de desarrollar los clasificadores, de que las voces son numéricamente distinguibles y se les puede aplicar algoritmos de clasificación para conseguir el objetivo principal de este trabajo, crear sistemas de detección de Edema de Reinke.

En la parte final del trabajo hemos partido de los resultados obtenidos en el sistema GMM con una función gaussiana para cada etiqueta, sanos y ER, ya que, de los tres sistemas tratados, es el que diríamos que es más sencillo. A partir de él, el objetivo era superar el porcentaje de precisión tanto en el sistema GMM con dos modelos distintos, como en la red neuronal y en ambos hemos conseguido cumplir el objetivo.

Comparando los resultados del sistema GMM con dos modelos y la Red Neuronal podemos llegar a la conclusión de que el primero se adapta mejor a los grupos de datos con los que contamos, posiblemente por la cantidad de datos disponibles. Debido a esto, podríamos afirmar que para grupos de datos limitados es más preciso recurrir a algoritmos de aprendizaje no supervisado tales como el GMM, ya que se adaptan rápidamente a los datos. En cambio, si se cuenta con grupos de datos de gran tamaño es mejor utilizar

algoritmos de aprendizaje profundo como las Redes Neuronales Multicapa, ya que son capaces de extraer características y patrones que otros algoritmos más sencillos no lo son.

Al fin y al cabo este trabajo tiene como objetivo principal el mismo que la propia definición básica de tecnología: *la resolución de problemas, y la creación de servicios que faciliten la adaptación al medio ambiente y la satisfacción de las necesidades esenciales y los deseos de la humanidad.*

La rama de la salud resulta esencial en el estudio de las capacidades de la tecnología, porque muchos desconocidos del campo pueden pensar (y con cierta razón) que los fines del desarrollo tecnológico se reducen a campos bélicos o publicitarios, a matar o envenenar a la gente, de manera directa o indirecta, pero causando un daño a la población.

Este trabajo supone una reivindicación, una necesidad de la tecnología en esa vuelta a la humanización inicial, a su verdadera definición, a su verdadera consumidora, la humanidad.

7. Bibliografía

A. Gonzalez et al. (2020). *CAMBIOS EN LOS PARÁMETROS OBJETIVOS Y SUBJETIVOS DE LA VOZ DE PACIENTES CON EDEMA DE REINKE TRAS TRATAMIENTO QUIRÚRGICO.*

al, Y. e. (1982). *Harmonics-to noise ratio as an index of the degree of hoarseness.*

Boersma et al. (1993). *Accurate short-term analysis of the fundamental frequency and the harmonic-to-noise ratio of a sample sound.*

Calvo, D. (20 de 07 de 2017). *diegocalvo.es*. Obtenido de https://www.diegocalvo.es/red-neuronal-convolucional/

D.Michaelis et al. (1996). *Glottal-to-Noise Excitation Ratio – a New Measure for Describing.* Göttingen, Germany.

Davis et al. (1980). *COMPARISON OF PARAMETRIC REPRESENTATIONS FOR MONOSYLLABIC WORD RECOGNITION IN.*

Godino et al. (2014). Cepstral peak prominence: A comprehensive analysis. En *Biomedical Signal Processing and Control.*

Halawa et al. (2012). *Estudio epidemiológico de pacientes con disfonías funcionales.* Otorrinolaringología.

Harar, P., Alonso-Hernandez, J., Mekyska, J., Galaz, Z., Burget, R., & Smekal, Z. (2017). *Voice Pathology Detection Using Deep Learning: a Preliminary Study.* Retrieved from https://ieeexplore.ieee.org/document/7985525

Healthtech. (06 de 06 de 2020). *HEALTH TECH SPAIN*. Obtenido de https://www.healthtechspain.es/telemedicina-espana-ejemplos-aplicaciones/

Hegde, S., Shetty, S., Rai, S., & Dodderi, T. (2018). A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. *journal of Voice.* Obtenido de https://www.jvoice.org/article/S0892-1997(18)30143-7/fulltext

Hillenbrand et al. (1994). Acoustic Correlates of Breathy Vocal. En *Joumal of Speech and Hearing Researc.*

Hospital, e. (06 de 06 de 2020). *http://www.elhospital.com/*. Obtenido de https://www.healthtechspain.es/telemedicina-espana-ejemplos-aplicaciones/

ISEP. (17 de 4 de 2017). *Instituto Superior de Estudios Psicológicos*. Obtenido de https://www.isep.es/actualidad-logopedia/guia-practica-rehabilitacion-del-edema-de-reinke/#:~:text=Edema%20de%20Reinke%3A%20Diagn%C3%B3stico%20y%20tratamiento&text=Es%20el%20resultado%20de%20un,la%20mucosa%20que%20lo%20cubre.

Kasuya et al. (1986). *Normalized noise energy as an acoustic measure to evaluate.* Utsonomiya.

Lyons, J. (16 de 8 de 2017). *pypi.org*. Obtenido de https://pypi.org/project/python_speech_features/

Mateos, C., & Jiménez, R. (2016). *seorl.* Obtenido de https://seorl.net/wp-content/uploads/2016/05/NP_Dia-Mundial-de-la-Voz.pdf

Merk et al. (1999). ACOUSTIC ASSESSMENT OF NEUROGENIC VOICE DISORDERS IN A CLINICAL SETTING. *Models and Analysis of Vocal Emissions for Biomedical Applications.* Firence.

Reyes Burneo, P. (2018). *http://www.pabloreyesotorrino.com/*. Obtenido de http://www.pabloreyesotorrino.com/voz-laringe/edema-reinke/

Teixeira et al. (2013). Vocal Acoustic Analysis - Jitter, Shimmer and HNR Parameters. *CENTERIS 2013 - Conference on ENTERprise Information Systems / HCIST 2013 - International.*

Teixeira, J. P., & Gonçalves, A. (2016). *Algorithm for jitter and shimmer measurements in pathologic voices.*

Teixeira, J. P., Odete Fernandes, P., & Alves, N. (2017). *Science Direct.* Obtenido de www.sciencedirect.com

University, S. (s.f.). *Saarbruecken Voice Database*. Obtenido de http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4

Wikipedia. (s.f.). Obtenido de https://en.wikipedia.org/wiki/Bayesian_information_criterion

Y. Maryn et al. (2009). Acoustic measurement of overall voice quality: a meta-analysis.

8. Anexos

8.1. ANEXO I: Encuestas ECV

Fecha de examen:.....
Realizada por.....
Nº Hª.....

DATOS PERSONALES

Apellidos: .....
Nombre:.....
Edad:.....Sexo .....
Personas con las que convive (indicar edades y relación familiar): .....
.....
Profesión:.....
Lugar habitual de residencia .....
Diagnóstico inicial:.....
.....

ASPECTOS RELACIONADOS CON LA SALUD EN GENERAL:

Tabaco u otras sustancias: Sí No
En la actualidad o anteriormente, cuánto fuma?
- 10 + 10 + 20 + 30 +40
Alcohol - consumo habitual.....
Café o té - consumo habitual - .....
Bebidas estimulantes - consumo habitual- .....
Las bebidas frías dañan su voz .....
Reflujo gastro-esofágico:.....
Trastornos psicoemocionales (estrés, depresión…).....
Infecciones de repetición: .....

ASPECTOS RELACIONADOS CON LA VOZ EN GENERAL:

Cuánto tiempo lleva con problemas de voz: .....

Desde que comenzó el trastorno:

- ha ido en aumento
- permanece igual
- es intermitente

En algún momento han desaparecido totalmente .....

¿En qué momento del día siente más dificultades o nota peor su voz?

- Por la mañana.
- Por la tarde.
- Por la noche.

¿En qué situación/es aparecen con más frecuencia sus molestias al hablar?

- En casa
- En el trabajo
- Con sus amistades
- Otras (indicar cuáles) .....

¿En qué ocasión aparecieron estos trastornos?

- Después de una gripe, tras un enfriamiento
- Después de un periodo de intenso trabajo o de gran fatiga
- Tras problemas profesionales o familiares
- Después de una intervención quirúrgica
- Tras un accidente
- Otras (indicar cuáles).....

¿Qué circunstancias aumentan su ronquera y fatiga'?

- Fumar mucho.....
- Ambientes muy ruidosos.....
- Dormir poco.....
- Hablar en público un tiempo continuado. .....
- Otras (indicar cuáles).....

Siendo 1 = nada/nunca/mal y 5 mucho/siempre/muy bien

PERCEPCIÓN VOCAL					
Valore la calidad de su voz	1	2	3	4	5
Empeora su voz a lo largo del día	1	2	3	4	5
Empeora su voz a lo largo de la semana	1	2	3	4	5

Mejora su voz si permanece en silencio	1	2	3	4	5

SENSACIONES					
Sequedad	1	2	3	4	5
Picor o irritación	1	2	3	4	5
Ardores	1	2	3	4	5
Punzadas	1	2	3	4	5
Molestias al tragar	1	2	3	4	5
Dolor	1	2	3	4	5
Sensación de tener algo extraño	1	2	3	4	5
Molestias, pesadez en la parte posterior del cuello y/o hombros	1	2	3	4	5
Ninguna sensación especial	1	2	3	4	5
Necesita carraspear cuando habla	1	2	3	4	5
Siente que le falta aire cuando habla	1	2	3	4	5
Siente fatiga general al hablar	1	2	3	4	5
Le salen “gallos” cuando habla	1	2	3	4	5
Pierde la voz momentáneamente	1	2	3	4	5
Pierde la voz definitivamente	1	2	3	4	5

INTENSIDAD, TONO Y TIMBRE					
Mi voz en el trabajo es generalmente fuerte	1	2	3	4	5
Mi voz fuera del trabajo es generalmente fuerte	1	2	3	4	5
Siento fatiga vocal en una conversación prolongada	1	2	3	4	5
Siento que a mi voz le falta potencia	1	2	3	4	5

Tengo dificultades para llamar a alguien distante	1	2	3	4	5
Tengo dificultades para hablar por teléfono	1	2	3	4	5
Suelo limitar el uso de mi voz	1	2	3	4	5
Nota su voz enronquecida	1	2	3	4	5
Nota su voz desagradable	1	2	3	4	5

CONDICIONES AMBIENTALES					
Tiempo de habla durante su jornada laboral	1	2	3	4	5
Habla al aire libre	1	2	3	4	5
Su lugar de trabajo tiene las condiciones acústicas adecuadas	1	2	3	4	5
Nivel de ruido exterior en el entorno laboral	1	2	3	4	5
Su domicilio tiene las condiciones acústicas adecuadas	1	2	3	4	5
Nivel de ruido exterior en su domicilio	1	2	3	4	5
Considera que grita al hablar	1	2	3	4	5
Deja descansar su voz a lo largo del día	1	2	3	4	5
Le resulta difícil hablar en un ambiente ruidoso	1	2	3	4	5
Sus problemas de voz se agravan con la sobrecarga de trabajo	1	2	3	4	5
Sus problemas de voz se agravan con actividades extraprofesionales (deporte, cantar, teatro…)	1	2	3	4	5

CONDICIONES PSICOEMOCIONALES					
Se considera una persona nerviosa	1	2	3	4	5
Siente que está en constante agitación	1	2	3	4	5
Siente que vive de manera estresada	1	2	3	4	5
Tiene problemas para conciliar el sueño	1	2	3	4	5
Siente que se irrita fácilmente	1	2	3	4	5
Sus problemas de voz se agravan con las preocupaciones o problemas	1	2	3	4	5

8.2. ANEXO II: Encuestas VHI

Marque con un círculo la respuesta que indique con qué frecuencia tiene la misma experiencia según la siguiente escala: **0 = nunca, 1 = casi nunca, 2 = algunas veces, 3 = casi siempre, 4 = siempre**

Parte I- F (Funcional)					
F.1. La gente me oye con dificultad debido a mi voz	**0**	**1**	**2**	**3**	**4**
F.2. La gente no me entiende en sitios ruidosos	**0**	**1**	**2**	**3**	**4**
F.3. Mi familia no me oye si la llamo desde el otro lado de la casa	**0**	**1**	**2**	**3**	**4**
F.4. Uso el teléfono menos de lo que desearía	**0**	**1**	**2**	**3**	**4**
F.5. Tiendo a evitar las tertulias debido a mi voz	**0**	**1**	**2**	**3**	**4**
F.6. Hablo menos con mis amigos, vecinos y familiares	**0**	**1**	**2**	**3**	**4**
F.7. La gente me pide que repita lo que les digo	**0**	**1**	**2**	**3**	**4**
F.8. Mis problemas con la voz alteran mi vida personal y social	**0**	**1**	**2**	**3**	**4**
F.9. Me siento desplazado de las conversaciones por mi voz	**0**	**1**	**2**	**3**	**4**
F.10. Mi problema con la voz afecta al rendimiento laboral	**0**	**1**	**2**	**3**	**4**
Parte II- P (Física)					
P.1. Me quedo sin aire al hablar	**0**	**1**	**2**	**3**	**4**
P-2. Mi voz suena distinto a lo largo del día	**0**	**1**	**2**	**3**	**4**
P.3. La gente me pregunta ¿Qué te pasa con la voz?	**0**	**1**	**2**	**3**	**4**
P.4. Mi voz suena cascada y seca	**0**	**1**	**2**	**3**	**4**
P.5. Siento que necesito tensar la garganta para producir la voz	**0**	**1**	**2**	**3**	**4**
P.6. La calidad de mi voz es impredecible	**0**	**1**	**2**	**3**	**4**
P.7. Trato de cambiar mi voz para que suene diferente	**0**	**1**	**2**	**3**	**4**
P.8. Hago bastante esfuerzo para hablar	**0**	**1**	**2**	**3**	**4**
P.9. Mi voz empeora por la tarde	**0**	**1**	**2**	**3**	**4**
P.10. Mi voz “se me acaba” a la mitad del habla	**0**	**1**	**2**	**3**	**4**
Parte III- E (Emocional)					
E.1. Me siento tenso al hablar con otros debido a mi voz	**0**	**1**	**2**	**3**	**4**
E.2. Las personas parecen irritadas por mi voz	**0**	**1**	**2**	**3**	**4**
E.3. Creo que la gente no comprende mi problema con la voz	**0**	**1**	**2**	**3**	**4**

E.4. Mi problema vocal me molesta	**0**	**1**	**2**	**3**	**4**
E.5. Salgo menos por mi problema de voz	**0**	**1**	**2**	**3**	**4**
E.6. Mi voz me hace sentir en desventaja	**0**	**1**	**2**	**3**	**4**
E.7. Me siento contrariado cuando me piden que repita lo dicho	**0**	**1**	**2**	**3**	**4**
E.8. Me siento avergonzado cuando me piden que repita lo dicho	**0**	**1**	**2**	**3**	**4**
E.9. Mi voz me hace sentir incompetente	**0**	**1**	**2**	**3**	**4**
E.10. Me avergüenza mi problema de la voz	**0**	**1**	**2**	**3**	**4**

8.3. ANEXO III: Tablas de estadísticas por vocal

En este anexo se muestran las estadísticas de los parámetros extraídas de los corpus de parámetros divididos por vocal pronunciada.

8.3.1. F0

F0 vocal A				
	MN	MR	FN	FR
count	129	24	276	219
mean	136,25	128,99	238,80	177,13
std	41,41	63,50	51,79	51,63
min	73,73	86,02	108,84	71,43
25%	106,67	99,61	207,79	136,75
50%	125,98	110,73	235,29	173,91
75%	156,89	136,48	261,02	207,79
max	314,68	405,6483	461,44	344,00

F0 vocal I				
	MN	MR	FN	FR
count	129	24	276	219
mean	144,82	127,76	245,44	190,45
std	49,42	25,62	50,88	53,02
min	73,39	88,89	112,24	77,48
25%	113,50	107,38	214,13	152,30
50%	133,33	121,82	242,42	188,26
75%	168,42	144,47	271,19	216,22
max	367,29	175,82	421,05	358,34

F0 vocal U				
	MN	MR	FN	FR
count	129	24	276	219
mean	142,91	128,75	246,97	186,27
std	47,17	30,55	50,46	56,31
min	59,42	66,39	109,04	73,39
25%	110,41	110,16	216,53	144,18
50%	128,66	118,97	239,76	180,89
75%	162,95	147,20	268,81	214,33
max	353,90	198,34	466,98	421,05

8.3.2. Jitter

Jitter vocal A				
	MN	MR	FN	FR
count	129	24	276	219
mean	1,04	2,20	1,31	3,04
std	1,19	2,07	1,09	2,56
min	0,25	0,43	0,32	0,30
25%	0,45	0,67	0,67	0,94
50%	0,64	0,93	0,95	2,17
75%	1,14	3,36	1,50	4,73
max	9,99	8,32	9,97	12,09

Jitter vocal I				
	MN	MR	FN	FR
count	129	24	276	219
mean	1,52	2,52	2,50	3,90
std	1,39	2,08	1,42	2,65
min	0,18	0,46	0,25	0,31
25%	0,62	1,48	1,47	1,91
50%	1,12	2,07	2,21	3,27
75%	2,00	2,53	3,32	5,68
max	8,86	10,33	6,99	13,86

Jitter vocal U				
	MN	MR	FN	FR
count	129	24	275	218
mean	1,16	2,37	1,43	2,79
std	1,11	3,53	0,85	2,40
min	0,23	0,42	0,37	0,40
25%	0,55	0,61	0,94	1,07
50%	0,85	0,86	1,24	1,85
75%	1,14	1,77	1,65	3,79
max	7,35	14,78	7,34	16,50

8.3.3. Shimmer

Shimmer vocal A				
	MN	MR	FN	FR
count	129	24	276	219
mean	4,40	27,96	6,53	14,48
std	3,98	74,30	4,77	23,93
min	0,73	0,49	0,88	0,46
25%	2,05	3,59	3,67	3,68
50%	3,41	4,88	5,33	7,51
75%	5,64	8,99	8,22	13,06
max	32,49	270,5036	40,52	201,91

Shimmer vocal I				
	MN	MR	FN	FR
count	129	24	276	219
mean	3,68	5,00	5,59	10,65
std	3,55	4,10	3,45	14,81
min	0,61	1,05	0,75	0,55
25%	1,88	2,10	3,13	3,27
50%	2,61	3,58	4,78	5,93
75%	3,94	5,70	7,13	12,00
max	29,77	15,05	20,70	100,56

Shimmer vocal U				
	MN	MR	FN	FR
count	129	24	275	218
mean	2,51	4,42	4,33	8,53
std	1,79	4,68	3,25	10,86
min	0,50	0,87	0,66	0,29
25%	1,38	1,92	2,23	2,34
50%	2,07	2,79	3,50	4,45
75%	2,99	4,19	5,21	9,58
max	11,60	20,31	20,05	70,70

8.3.4. HNR

HNR vocal A				
	MN	MR	FN	FR
count	129	24	276	219
mean	23,33	17,14	24,44	12,67
std	4,64	7,50	4,39	17,11
min	-3,51	2,03	-7,77	-116,07
25%	21,45	14,32	22,76	9,44
50%	23,66	19,44	24,92	17,42
75%	26,11	21,80	26,81	21,62
max	31,45	27,36	34,08	29,59

HNR vocal I				
	MN	MR	FN	FR
count	129	24	276	219
mean	25,81	22,04	27,77	18,28
std	5,02	6,69	4,58	12,53
min	-9,97	8,49	4,74	-39,93
25%	23,36	17,82	24,96	15,27
50%	26,13	25,74	27,86	21,19
75%	28,71	26,77	30,65	26,24
max	37,80	31,72	41,88	36,87

HNR vocal U				
	MN	MR	FN	FR
count	129	24	275	218
mean	28,30	24,38	28,61	19,33
std	4,88	10,35	5,23	12,21
min	-2,97	0,02	6,13	-39,83
25%	26,17	17,80	26,11	13,62
50%	28,52	27,86	29,46	22,42
75%	30,94	31,24	32,25	27,57
max	37,92	36,85	39,70	40,34

8.3.5. NHR

NHR vocal A				
	MN	MR	FN	FR
count	129	24	276	219
mean	-30,80	-28,66	-30,79	-29,76
std	4,55	5,30	4,74	5,80
min	-41,88	-38,10	-43,00	-41,75
25%	-33,42	-32,46	-34,30	-33,71
50%	-30,64	-29,17	-31,24	-29,91
75%	-28,10	-26,70	-27,64	-26,61
max	-19,17	-17,00	-17,12	-7,22

NHR vocal I				
	MN	MR	FN	FR
count	129	24	276	219
mean	-24,05	-23,09	-28,03	-26,27
std	6,50	7,57	5,54	6,26
min	-44,29	-39,37	-43,09	-41,29
25%	-29,10	-26,91	-31,44	-29,85
50%	-24,11	-21,67	-28,25	-26,69
75%	-19,17	-18,66	-24,87	-22,87
max	-8,35	-11,47	-9,57	-6,45

NHR vocal U				
	MN	MR	FN	FR
count	129	24	275	218
mean	-45,20	-43,33	-45,47	-44,74
std	8,76	6,21	7,94	9,93
min	-85,65	-52,31	-88,82	-90,00
25%	-47,44	-47,92	-46,48	-46,90
50%	-44,40	-44,89	-44,62	-44,24
75%	-40,66	-38,71	-42,43	-40,86
max	-26,32	-30,31	-32,01	-6,71

8.3.6. GNE

GNE vocal A				
	MN	MR	FN	FR
count	129	24	276	219
mean	0,89	0,79	0,81	0,74
std	0,10	0,11	0,14	0,15
min	0,44	0,61	0,35	0,32
25%	0,85	0,71	0,73	0,65
50%	0,93	0,80	0,86	0,77
75%	0,96	0,88	0,92	0,86
max	0,99	0,94	0,99	0,97

GNE vocal I				
	MN	MR	FN	FR
count	129	24	276	219
mean	0,87	0,82	0,79	0,74
std	0,10	0,11	0,15	0,16
min	0,52	0,59	0,24	0,24
25%	0,83	0,70	0,72	0,63
50%	0,91	0,86	0,84	0,75
75%	0,95	0,92	0,91	0,87
max	0,99	0,97	0,97	0,98

GNE vocal U				
	MN	MR	FN	FR
count	129	24	275	218
mean	0,72	0,68	0,66	0,62
std	0,16	0,18	0,15	0,16
min	0,28	0,29	0,30	0,22
25%	0,59	0,57	0,55	0,49
50%	0,75	0,74	0,66	0,59
75%	0,86	0,82	0,77	0,74
max	0,96	0,90	0,96	0,96

8.3.7. NNE

NNE vocal A				
	MN	MR	FN	FR
count	129	24	276	219
mean	-5,39	-3,77	-4,40	-3,76
std	3,31	2,67	3,24	2,32
min	-18,16	-9,36	-16,28	-12,66
25%	-7,49	-4,95	-6,05	-4,82
50%	-4,86	-3,26	-3,88	-3,15
75%	-3,27	-2,14	-2,04	-2,20
max	3,84	-0,10	3,02	0,01

NNE vocal I				
	MN	MR	FN	FR
count	129	24	276	219
mean	-4,38	-3,66	-4,37	-3,05
std	1,93	1,86	1,78	1,78
min	-9,08	-8,28	-11,80	-11,21
25%	-5,62	-4,84	-5,42	-3,57
50%	-3,84	-3,05	-4,20	-2,54
75%	-3,13	-2,21	-3,17	-1,86
max	0,25	-1,37	2,34	-0,18

NNE vocal U				
	MN	MR	FN	FR
count	129	24	275	218
mean	-2,04	-2,19	0,93	-0,25
std	2,33	1,91	3,18	1,96
min	-7,68	-7,20	-5,65	-7,48
25%	-3,18	-3,41	-1,26	-1,32
50%	-1,95	-1,79	0,29	-0,48
75%	-1,28	-1,42	2,39	0,50
max	13,31	2,74	14,41	7,38

8.3.8. CPP

CPP vocal A				
	MN	MR	FN	FR
count	129	24	276	219
mean	20,5	15,9	21,4	14,9
std	3,7	4,0	3,1	5,2
min	10,6	8,3	10,7	3,7
25%	18,4	12,4	19,6	10,6
50%	20,9	16,6	21,7	15,5
75%	22,7	19,0	23,8	18,9
max	28,0	22,5	28,2	25,3

CPP vocal I				
	MN	MR	FN	FR
count	129	24	276	219
mean	19,7	16,3	20,2	14,3
std	3,3	4,6	3,1	5,1
min	11,6	8,8	8,0	2,6
25%	17,8	12,5	18,2	11,0
50%	20,3	15,6	20,3	14,0
75%	22,0	19,6	22,6	18,3
max	27,5	26,5	27,1	28,8

CPP vocal U				
	MN	MR	FN	FR
count	129	24	275	218
mean	18,56	16,75	18,52	14,84
std	2,98	3,17	2,58	4,22
min	9,95	9,88	8,57	3,60
25%	16,47	15,06	16,73	12,57
50%	18,56	17,47	18,33	15,28
75%	20,63	18,90	20,19	17,77
max	26,36	21,75	25,20	25,32

8.4. ANEXO IV: Tablas de resultados GMM

En este anexo, se muestran los resultados de eficiencia del sistema GMM variando el número de funciones gaussianas que componen cada modelo.

GMM 2	DF_A	DF_I	DF_U	DF
media	84,62	93,79	85,69	90,03
std	0,73	1,77	2,99	2,37
min	83,08	89,23	79,49	85,10
max	85,64	96,92	90,26	92,12

GMM 3	DF_A	DF_I	DF_U	DF
media	84,77	92,15	84,00	90,98
std	3,14	3,14	4,74	1,80
min	81,03	85,64	75,90	85,96
max	91,79	98,46	92,82	92,81

GMM 4	DF_A	DF_I	DF_U	DF
media	88,87	91,13	90,31	90,92
std	5,10	1,49	3,35	1,09
min	78,97	88,21	85,13	87,84
max	97,44	92,82	95,38	91,78

GMM 5	DF_A	DF_I	DF_U	DF
media	86,62	88,62	86,92	90,29
std	5,04	1,55	2,41	1,48
min	77,95	86,67	84,10	87,84
max	93,85	91,28	91,28	92,29

GMM 6	DF_A	DF_I	DF_U	DF
media	87,23	85,85	79,79	91,11
std	3,76	5,18	5,33	1,71
min	80,00	77,95	69,23	88,36
max	91,79	96,41	87,69	94,35

GMM 7	DF_A	DF_I	DF_U	DF
media	75,79	84,92	88,26	94,02
std	6,87	4,85	3,57	1,88
min	65,13	74,36	83,08	90,75
max	87,69	91,28	95,38	96,23

GMM 8	DF_A	DF_I	DF_U	DF
media	80,41	80,41	79,90	93,53
std	8,12	6,60	7,63	1,84
min	65,13	70,26	68,72	89,73
max	92,82	91,79	95,90	96,06

GMM 9	DF_A	DF_I	DF_U	DF
media	75,44	83,38	80,05	91,61
std	8,66	6,54	8,98	2,34
min	65,64	66,67	64,62	86,47
max	92,31	90,26	97,95	94,69

GMM 10	DF_A	DF_I	DF_U	DF
media	76,82051	73,02564	71,4359	89,7774
std	7,071719	7,062043	9,267874	1,171538
min	66,15385	64,61538	63,58974	86,9863
max	87,69231	83,07692	87,69231	91,26712