Trabajo Fin de Grado

Análisis automático de la señal de voz para el diagnóstico clínico y la valoración de trastornos en el habla
Automatic analysis of the voice signal and evaluation of disorders in the speech

Autora

Inés Pérez Serrano

Director

Eduardo Lleida Solano

Escuela de Ingeniería y Arquitectura
2018

TRABAJOS DE FIN DE GRADO / FIN DE MÁSTER

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe acompañar al Trabajo Fin de Grado (TFG)/Trabajo Fin de Máster (TFM) cuando sea depositado para su evaluación).

D./Dª. Inés Pérez Serrano,

con nº de DNI 17456447E en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza,

Declaro que el presente Trabajo de Fin de (Grado/Máster) Grado, (Título del Trabajo) Análisis automático de la señal de voz para el diagnóstico clínico y la valoración de trastornos en el habla.

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, 20 de noviembre de 2018

Fdo: Inés Pérez Serrano

Agradecimientos

A Eduardo Lleida por su paciencia en este proyecto.

A mis padres, por su apoyo y porque gracias a ellos, hoy estoy donde estoy.

A Ana y a Ramón, por soportarme y cuidarme.

A todas esas personas que he conocido en estos años en la Escuela, porque habéis pasado de ser compañeros a ser amigos.

Y a todos aquellos que hicieron que mi tema tabú, este proyecto, dejara de ser tabú.

ANÁLISIS AUTOMÁTICO DE LA SEÑAL DE VOZ PARA EL DIAGNÓSTICO CLÍNICO Y LA VALORACIÓN DE TRASTORNOS EN EL HABLA

Resumen

La necesidad del ser humano de relacionarse con el entorno que le rodea hace de la comunicación hablada una habilidad prácticamente imprescindible. Cualquier perturbación en la capacidad del habla puede acarrear efectos negativos en el bienestar de una persona. Ante esta circunstancia, un diagnóstico temprano puede mejorar sustancialmente la vida del paciente. Una herramienta informática que ofrezca unas guías totalmente objetivas sobre los posibles trastornos en el habla que presente una persona, puede ayudar significativamente a los profesionales que se dedican a este campo.

Este proyecto se estructura en las dos partes que indica el título. Por un lado, el análisis clínico se encarga de extraer una serie de parámetros de la señal de voz, que aportan información del estado de la voz en distintos ámbitos. Se analiza cómo evoluciona la amplitud de la señal, si presenta saltos; es decir, si el paciente es capaz de emitir una vocal durante unos segundos de forma continua. También se analiza el ruido presente de distintas formas: relación de los harmónicos frente a ruido, energía glotal de ruido, etc.

La segunda parte, haciendo uso de redes neuronales, valora si los datos extraídos de la señal de voz pueden asociarse a una patología o patologías concretas o por el contrario no hay apreciación de patologías en la voz. En concreto, se analiza un total de 71 patologías extraídas de la base de datos Saarbrüecken Voice Database. Además de voces patológicas, la base de datos presenta también voces de personas sanas. Aunque la base de datos provee grabaciones de las vocales /a/, /i/ y /u/, solo se han utilizado las relativas a la vocal /a/.

Finalmente, el análisis automático de las grabaciones de la base de datos y su posterior procesado con la red neuronal, ofrece una probabilidad de detección de patología en torno al 72%. Esta probabilidad sube hasta el 76-77% en caso de intentar detectar menor número de patologías, ya que en la muestra inicial hay patologías con escasa representación.

Sumario

Índice de figuras

Índice de tablas

Índice de siglas

ANN: Artificial Neural Network.
DNN: Deep Neural Networks, 2
DUV: Degree of Voiceless.
DVB: Degree of Voice Breaks, 10
GMM: Gaussians Mixture Model.
GNE: Glottal to Noise Excitation Ratio.
HNR: Harmonic to Noise Ratio.
MFCC: Mel-Frequency Cepstral Coefficients.
NHR: Noise to Harmonic Ratio.
NNE: Normalized Noise Energy.
PFR: Phnatory Fundamental Frequency Range.
RAP: Relative Average Perturbation.
sAPQ: smoothed Amplitude Perturbation Quotient.
SPI: Soft Phonation Index.
sPPQ: smoothed Pitch Period Perturbation Quotient.
std: desviación estándar.
SVD: Saarbruecken Voice Database.
vAm: variación de Amplitud.
vFo: varación de la Frecuencia Fundamental.
VTI: Voice Turbulence Index.
ZCR: Zero Crossing Rate.

1. Introducción

1.1 Motivación

El ser humano, como ser social que es, necesita comunicarse con el medio que le rodea. Una de las principales formas de comunicación es la oral y, por tanto, una afección o pérdida de la capacidad del habla puede afectar al bienestar de una persona.

La variedad y las peculiaridades de los trastornos del habla, así como sus causas, los estudia la Logopedia. El diagnóstico de cualquier patología del habla lo realizan profesionales muy entrenados que, escuchando la voz del paciente, son capaces de discernir qué posibles patologías pueden tener. Sin embargo, el proceso puede ser subjetivo, ya que depende tanto de la colaboración del paciente como de la experiencia del profesional.

Con este trabajo se pretende ofrecer una herramienta que ayude a logopedas a quitar el sesgo subjetivo, ofreciendo unas guías para el diagnóstico basadas en el análisis de la señal de voz del paciente, combinada con metodologías de aprendizaje automático.

1.2 Estado del arte

El presente proyecto se ha desarrollado dentro del Área de Teoría de la Señal y Comunicaciones de la Escuela de Ingeniería y Arquitectura de la Universidad de Zaragoza, bajo la tutela del grupo de investigación VivoLab. La primera aproximación teórica viene de la mano de varios trabajos desarrollados por este grupo. El primero de ellos, *Automatic GRBAS Rating for Voice Quality Assessment Using Multidimensional* [1], utiliza la escala GRBAS para realizar una evaluación de la calidad de la señal con fines diagnósticos; además, emplea la misma base de datos empleada en este estudio, *Saarbruecken Voice Database*(SVD).

Otro estudio [2] analiza la base de datos SVD, extrayendo parámetros relacionados con el ruido presente en la señal y parámetros espectrales, para después clasificarlos con un clasificador GMM (Modelo de Mezclas de Gaussianas) en busca de patologías, con una precisión en torno al 79%.

Fuera de este grupo de investigación, otros trabajos inspiradores han sido [3] y [4]. El primero realiza una clasificación binaria para distinguir si una muestra de voz es sana o patológica, diferenciando entre voces masculinas y femeninas. Este estudio consigue mejores resultados en la base de datos de mujeres, con un rendimiento del 95% frente al 88% obtenido con las voces masculinas. El segundo, analiza una serie de parámetros relacionados con la amplitud, la frecuencia de la señal y la relación HNR y concluye que existe la posibilidad de realizar la clasificación de forma remota, ya que utiliza una base de datos compuesta por grabaciones recopiladas durante conversaciones telefónicas.

Por último, se buscaron otros estudios que hubiesen utilizado la base de datos SVD, entre ellos [5], que hace una recopilación de estudios previos además de aportar sus propias conclusiones. En dicho estudio, utilizan *Deep Neural Networks* (DNN) y, en vez de extraer parámetros acústicos de la señal, los parámetros de entrada de la red son vectores de la señal enventanada cada 64ms. El coste computacional de esta aproximación es alto y detecta si una voz es patológica con una probabilidad del 71,36%. El resto de los estudios referenciados en este último documento reflejan precisiones de entre el 72% hasta el 100%, utilizando distintas características y clasificadores. El único de estos estudios que utiliza redes neuronales artificiales emplea una submuestra de la red neuronal, que contiene 4 de las 71 patologías, a saber, laringitis crónica, quiste, edema de Reinke y disfonía espasmódica. Los parámetros de entrada a la red neuronal son los MFCC (*Mel-Frequency Cepstral Coefficients*) y sus primera y segunda derivadas. Consigue una clasificación correcta en el 87,82% de los casos.

1.3 Objetivos

El principal objetivo del presente trabajo consiste en proporcionar una herramienta informática capaz de decidir si una persona presenta un trastorno del habla o no y, si lo padece, estimar cuál o cuáles serían los más probables.

Esta herramienta analiza varios aspectos de la señal de voz, como son los referidos a amplitud, frecuencia, presencia de ruido o contenido espectral. Una vez calculados estos parámetros, se pasan a un clasificador que indicará si la voz analizada presenta una patología o no. En el caso de que se detecte que la voz es patológica, listará las cuatro patologías más probables presentes en dicha señal.

Por otra parte, las voces femeninas presentan frecuencias de pitch (vibración de las cuerdas vocales) más altas que las masculinas, y una forma de mejorar los procesos que incluyen aprendizaje automático es procesar voces masculinas y femeninas de forma diferenciada; por tanto, la herramienta permitirá elegir qué tipo de voz se desea analizar.

Por último, la aplicación ofrece una representación gráfica de todos los parámetros analizados, contrastados con los parámetros medios referentes a una persona sin patologías.

1.4 Herramientas

Como paso previo a la programación del proyecto, ha habido un proceso para analizar qué herramientas serían las más adecuadas para llevarlo a cabo. Como colofón al análisis automático de la voz se quería desarrollar una prueba de concepto, en forma de aplicación que permita procesar nuevas señales de voz.

En un inicio, se había pensado en que la aplicación se ejecutase en un terminal móvil con Android. Java es el lenguaje más ampliamente utilizado para desarrollar en Android; sin embargo, la función principal de la aplicación es procesar señales de audio y su tratamiento en Java es bastante tedioso al carecer en su sintaxis de

métodos matriciales que optimizan un procesado tan básico como puede ser calcular una transformada de Fourier.

Descartado el uso de Java, había que investigar sobre otros lenguajes. Matlab es el idóneo para tratar señales, pero su puesta en marcha en Android podría resultar farragosa; por tanto, quedó descartado. El siguiente lenguaje analizado fue Python. Python proporciona numerosas librerías como Pandas o Numpy que permiten hacer infinidad de cálculos, como transformadas o convoluciones, de una forma muy eficiente. La siguiente pregunta fue, ¿cómo interactúa Python con Android?

Existe un *framework*, llamado Kivy [6], que permite ejecutar código Python en varias plataformas: Android, Windows, Linux, iOs. Es decir, con esta opción se permite realizar un procesado de señal eficiente utilizando Python y además la aplicación podría ser multiplataforma. Esto quiere decir se puede conseguir una aplicación de escritorio, cuyo testeo va a ser muy sencillo y que en un futuro se podría exportar a Android.

Otra de las bondades de Python es la cantidad de documentación, bibliotecas y *frameworks* disponibles para aprendizaje automático. En este proyecto se utiliza el *framework* de TensorFlow [7] y las bibliotecas implicadas en este apartado son scikit-learn [8] y Keras [9].

Para elegir el entorno de programación, había que valorar qué herramientas nos proporcionan la creación de entornos virtuales, para facilitar la instalación de las librerías necesarias para Python. Las dos opciones más comunes son Conda [10] y Pycharm [11]. Ambas están disponibles para Windows y Linux pero, por comodidad y facilidad de uso, se ha optado por utilizar PyCharm Community Edition en Windows.

Además, para poder realizar un control de versiones y asegurar el proyecto frente a posibles imprevistos, se ha utilizado GitLab en su versión privada.

Por último, aunque la aplicación haya sido desarrollada puramente en Python, algunas de las funcionalidades de procesado de señal se han desarrollado anteriormente en Matlab y así se han podido comparar los resultados obtenidos con los dos lenguajes.

1.5 Organización de la memoria

La memoria del trabajo fin de grado se ha dividido en 6 capítulos:

- Capítulo 1: Introducción

- Capítulo 2: Metodología
  Este capítulo se divide en varias secciones que abordan todo el proceso creativo. Parte de un esquema que sitúa al lector en el proyecto. La primera sección está dedicada a la base de datos elegida, además de como se ha utilizado en el proyecto.

La siguiente sección describe el análisis automático de la señal de voz y analiza uno a uno los parámetros que forman parte del procesado de señal que se pasará luego a la red neuronal. Dichos parámetros se pueden englobar en 5 grandes áreas: amplitud, frecuencia, contenido espectral, ruido y cambios de registro tonal.

A continuación, viene la sección de valoración de trastornos del habla, en la que se analiza la red neuronal empleada y como se ha llegado hasta ella.

Para finalizar este capítulo, se encuentra la sección de aplicación, donde se detalla el funcionamiento de la herramienta final. Se incluyen imágenes y el diagrama de flujo del proceso, desde que se inicia la aplicación hasta que muestra el resultado final por pantalla.

- Capítulo 3: Resultados
  En esta sección se analizan los resultados, tanto de la aplicación en sí, como los obtenidos por las redes neuronales.

- Capítulo 4: Conclusiones
  Aquí se detallan las conclusiones que saca la autora de la memoria de todo el trabajo detallado.

- Capítulo 5: Líneas futuras
  La herramienta ofrece unas posibilidades inalcanzables en el presente trabajo y por tanto se ofrecen futuras líneas que se podrían desarrollar en el futuro.

- Capítulo 6: Bibliografía
  Como el nombre indica, en esta sección se muestra toda la bibliografía utilizada en el proyecto.

2. Metodología

Como se ha mencionado en la introducción, el objetivo de este trabajo consiste en crear una herramienta que permita determinar si una voz presenta una patología o no. El proceso implicado tiene tres partes claves, a saber, base de datos utilizada, análisis automático de la señal de voz y la valoración de trastornos del habla. En la figura 2.1 se muestra un diagrama de bloques que relaciona estas partes con su función dentro del proyecto.

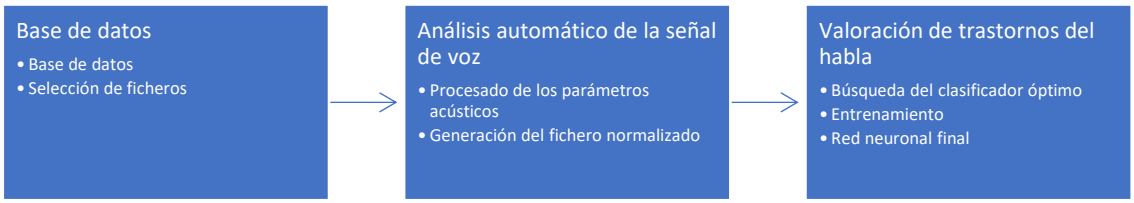

Figura 2.1: Diagrama de bloques de las funciones del proyecto

La base de datos estará formada por registros de voz tanto patológica como sana. En el caso de los registros de voz patológica deberá estar validada por un profesional indicando la patología o patologías que contiene. La base de datos se utilizará para el entrenamiento del clasificador y para realizar un test de prestaciones del sistema. Las patologías presentes en la base de datos nos definirán los tipos de trastornos que detectará nuestro sistema.

De la señal de voz se extraerán un conjunto de parámetros que serán útiles para la determinación de los trastornos de la voz; para ello se utilizarán técnicas clásicas de procesado digital de la señal. Por último, la valoración de los trastornos de la voz se realizará utilizando una red neuronal entrenada con la base de datos y a la que se le aplicarán los parámetros de la voz extraídos en la fase de análisis y cuya salida será la probabilidad de padecer cada una de las patologías definidas.

Por último, en la sección de aplicación se muestra la herramienta construida y la importancia que tienen los bloques anteriores en ella.

2.1 Base de datos

La base de datos utilizada en el presente trabajo es conocida como Saarbrüecken Voice Database [12], desarrollada por el Institut für Phonetik de la Universidad de Saarlandes. Se ha elegido esta base de datos por la gran variedad de patologías que recoge, además de su fácil acceso, ya que está publicada en la página web de la citada universidad de forma gratuita.

La base de datos recoge para cada orador una serie de grabaciones: las vocales /a/, / i/, /u/ con tono neutro, alto, bajo y alto-bajo-alto, además de una frase en alemán: ‘Guten Morgen, wie geht es Ihnen’, que significa ‘Buenos días, ¿cómo estás?’

Para proceder con la descarga de los ficheros se han de rellenar dos formularios. En el primero de ellos se puede filtrar por sexo, edad, patología o número de sesión. En el siguiente, se puede seleccionar el formato de los ficheros de la señal de voz y del EGG, además de las vocales mencionadas anteriormente.

En total, hay 1163 registros de voces femeninas, de las cuales 727 están clasificadas como patológicas. Respecto a las voces masculinas, 629 aparecen marcadas con al menos una patología de un total de 1062 registros. En cuanto a la parte técnica, las grabaciones están muestreadas a 50kHz con una resolución de 16-bit.

Para el desarrollo del presente trabajo se ha mantenido la separación entre voces femeninas y masculinas a la hora de valorar los posibles trastornos del habla, ya que hay estudios [3] que indican que esta diferenciación podría ser crítica. Por otra parte, no todas las patologías son igual de comunes y, por tanto, la base de datos no es homogénea en cuanto al número de grabaciones por patología. La base de datos tampoco es homogénea en el sentido de que el número de registros de voces sanas es inferior al registro total de voces patológicas. Además, hay patologías solo presentes en la base de datos de mujeres y viceversa.

Por ello, en la parte de valoración de trastornos del habla, hay que hacer una selección de la distribución de registros que se van a utilizar para evitar en la medida de lo posible sesgos en la red neuronal. Se van a utilizar dos aproximaciones; la primera de ellas cuenta con un 50% de voces sanas y un 50% patológicas, habiendo al menos una muestra de cada patología en cualquiera de los tonos, alto, bajo o normal. En la segunda, se emplean la totalidad de voces sanas y el 80% de las voces patológicas, asegurando de igual manera la existencia de al menos una patología en el conjunto de muestras. Este proceso se aplica por igual a las voces femeninas y masculinas.

Respecto a la distribución de las patologías, se ha mencionado que no es homogénea. En la figura 2.2 se muestra la distribución de patologías en hombres. Se puede ver que las patologías con más registros son laringitis (83 grabaciones), parálisis recurrente (*Rekurrensparese, 74*), y *Kontaktpachydermie (69)*. En cambio, hiperastenia, condroma y rinofonía cerrada solo presentan un registro.

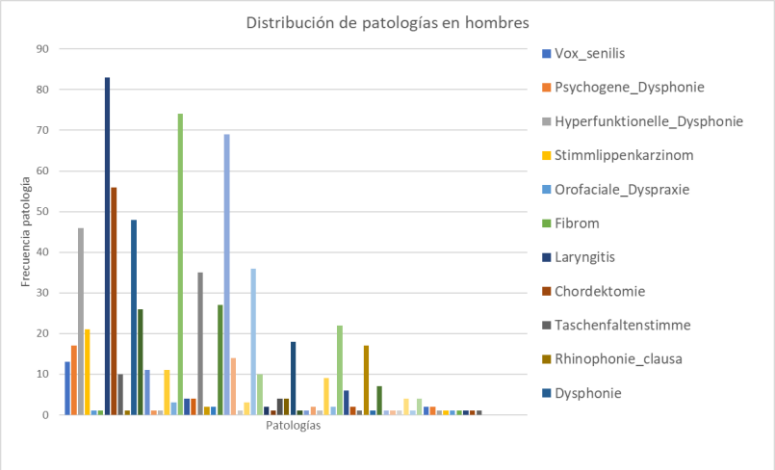

Figura 2.2 Distribución de patologías en hombres

En la figura 2.3 se muestra la distribución de patologías en mujeres. En el registro de voces femeninas, la patología más común es la disfonía hiperfuncional (167 muestras), seguida por la parálisis recurrente. Entre las patologías que solo encontramos una muestra están el fibroma, la diplofonía o la monocorditis.

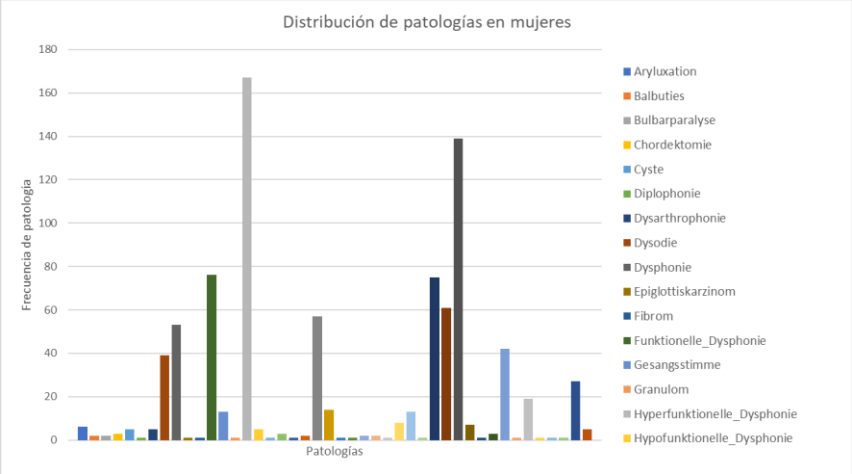

Figura 2.3 Distribución de patologías en mujeres

Tanto para voces patológicas como para las no patológicas, se ha seleccionado la vocal /a/ en sus tontos normal, alto y bajo para la investigación, porque la capacidad de sostener esta vocal durante unos segundos reporta información suficiente de la voz como para valorar trastornos del habla [13].

Por último, se han eliminado de la base de datos todas aquellas grabaciones de menores de 18 años, ya que en general las voces infantiles presentan unas frecuencias de pitch más altas y la muestra era poco representativa para considerar un grupo aparte. La relación completa de las patologías y su traducción al castellano se encuentra en el Anexo A.

2.2 Análisis automático de la señal de voz

Como se ha mencionado en la introducción, un logopeda puede detectar si una voz presenta algún desorden analizando los sonidos que emite el paciente sin necesidad de recurrir a instrumental técnico. En este trabajo se pretende que una máquina simule el proceso por el cual un logopeda puede discernir si un paciente presenta una patología o no. Para ello, lo primero que se hará será analizar la señal de voz y extraer una serie de parámetros. Estos parámetros se pueden englobar en cinco áreas: amplitud, frecuencia, contenido espectral, ruido y cambios de registro tonal o *voice breaks*. En las próximas páginas se detallarán todos los parámetros, así como una breve descripción sobre ellos.

<u>Parámetros de amplitud:</u>

Los parámetros que se calculan en este grupo dependen fundamentalmente de la diferencia entre la amplitud máxima y mínima de cada periodo de la señal que, para facilitar la comprensión de la memoria, llamaremos a este factor *A*. En la práctica, *A*

será el vector de *n* muestras que se pase a cada una de las funciones relacionadas con la amplitud de la muestra.

- Shimmer: es una medida que indica como varía la amplitud de la señal periodo a periodo. De este parámetro se ofrecen dos medidas, una en porcentaje y otra en decibelios. [4] [14]

$$Shimmer = \sum \frac{|A_i - A_{i+1}|}{\bar{A}} \quad [1]$$

$$\mathrm{Shimmer(dB)} = 20 * \log\left(\frac{1}{n-1}\sum \frac{|A_i - A_{i+1}|}{\bar{A}}\right) \quad [2]$$

Donde A_i es la amplitud del periodo actual y $\bar{A}$, el valor medio del vector de amplitudes.

- sAPQ: de sus siglas en inglés *smoothed Amplitude Perturbation Quotient*. Este parámetro calcula la variabilidad a largo plazo de la variación entre el máximo y mínimo de amplitud. Puede calcularse con varios factores de *smooth* o *sf;* en este caso, se ha utilizado con los factores *sf* de 5 y de 55 periodos. Al igual que en las ecuaciones anteriores, A_i denota la amplitud del periodo actual y $\bar{A}$, el valor medio de todas las amplitudes. [4] [14]

$$\mathrm{sAPQ} = \frac{\frac{1}{n-sf-1}\sum_{i=\frac{sf}{2}}^{n-\frac{sf}{2}}\left|\sum_{j=i-\frac{sf}{2}}^{i+\frac{sf}{2}}\frac{A_j}{sf} - A_i\right|}{\bar{A}} \quad [3]$$

- vAm: esta medida representa la desviación estándar del vector de amplitudes A. [4] [14]

<u>Parámetros frecuenciales:</u>

Al igual que en el caso anterior, la mayoría de las medidas dependen de dos vectores T y F, ambos de longitud n muestras, que contienen la duración de cada periodo de la señal y la frecuencia fundamental de cada periodo respectivamente.

- Jitter: expresa la variabilidad del pitch entre periodos. La variación entre dos periodos consecutivos viene dada por la duración del periodo actual T_i y el siguiente T_{i+1}. Devuelve el parámetro en segundos y en porcentaje. [4] [14]

$$\mathrm{jitta} = \sum \frac{|T_i - T_{i+1}|}{n-1} \quad [4]$$

- RAP: del inglés, *Relative Average Perturbation*, es un parámetro muy parecido al sAPQ de los parámetros amplitud. Calcula la variación del periodo de pitch con un factor de *smooth* de 3 periodos. [4] [14]
- sPPQ: es el mismo parámetro que en el caso anterior, pero en este caso con un factor de *smooth* de 5 y 55 periodos. [4] [14]

$$sPPQ = \frac{\frac{1}{n - sf - 1}\sum_{i=\frac{sf}{2}}^{n-\frac{sf}{x}}\left|\sum_{j=i-\frac{sf}{2}}^{i+\frac{sf}{2}}\frac{T_j}{sf} - T_i\right|}{\bar{T}} \quad [5]$$

Donde sf es el factor de smooth y T el vector de periodos T_i, con media $\bar{T}$ y longitud n.

- std: calcula la desviación estándar del vector F, es decir, la desviación estándar del de la frecuencia fundamental en cada periodo. [4] [14]
- vFo: calcula la desviación estándar relativa del vector F. [4] [14]
- PFR: de sus siglas en inglés, *Phonatory Fundamental Frequency Range*, es decir, el rango de frecuencias fundamentales en semi-tonos. La frecuencia máxima viene dada por $\mathrm{F0max}$ y la mínima por $\mathrm{F0min}$. [4] [14]

$$PFR = 12 * log_2 \frac{F0max}{F0min} \quad [6]$$

<u>Parámetros de envolvente espectral:</u>
En este apartado se calculan los MFCC, que son los coeficientes cepstrum utilizando la escala de frecuencias Mel (figura 2.4). Este es el único parámetro que no se ha implementado desde cero en Python, ya que, al ser un parámetro ampliamente utilizado, hay numerosas librerías que lo proveen, como *Python_speech_features* [15]. El resto de los parámetros devuelven un único resultado numérico, sin embargo, los MFCC son una matriz de tamaño NxM, donde M es el número de coeficientes y N el número de subtramas de la señal.

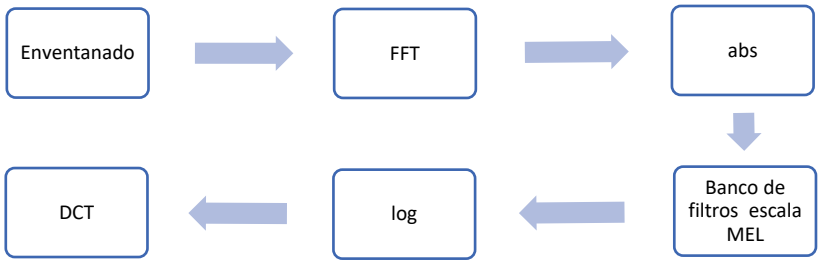

Figura 2.4 Extracción de MFCC

Dicha librería permite pasarle diferentes parámetros de entrada. En este caso se han calculado utilizando un tamaño de ventana de 40ms, con un solape entre tramas de 20ms. El número de coeficientes calculados por trama son 15 y la Transformada de Fourier tendrá 2048 puntos. [15] [16]

<u>Parámetros de *cambios de registro tonal*</u>:
Los parámetros incluidos en este apartado miden la capacidad de mantener la fonación en el tiempo, ya que en ocasiones las voces patológicas no pueden conseguirlo. Esta capacidad se va a medir mediante dos parámetros: *Degree of Voice*

Breaks (DVB) y *Degree of Voiceless*(DUV). El segundo, depende de un tercer parámetro llamado ZCR o *Zero Crossing Rate*.

- DVB: se define como la relación entre la duración de las áreas que presentan un '*break*' o salto respecto a la duración total de la señal. Para calcular si una señal presenta '*breaks*', compararemos si las distancias entre pulsos consecutivos superan un límite máximo. Este límite se define como $T_h = \frac{1.25}{pitch}$. En voces sanas, este parámetro idealmente sería cero. [14] [17]

- DUV: este parámetro calcula qué áreas de la señal se pueden marcar como '*unvoiced*', es decir, aquellas en las que no se puede detectar una frecuencia fundamental. Para diferenciar entre zonas sonoras y sin sonido, en esta aproximación se utiliza la energía de la señal y el parámetro mencionado anteriormente, ZCR. [18] [19] [20]

  La tasa de cruces por cero o ZCR [21], ecuación 7, indica las veces que la señal cambia de signo. En zonas en las que no hay sonido o en las que hay un salto de voz, esta tasa se incrementa notablemente. El umbral establecido que marca si un fragmento se considera sordo o sonoro está en 2500 cruces/segundo. En cambio, la energía de la señal es más elevada en zonas sonoras que en no sonoras. Para calcular este parámetro, se emplea la señal de voz *s* de N muestras, y la función signo, *sgn.*

$$ZCR = \frac{1}{N}\sum_{n=-\infty}^{\infty}\frac{1}{2}\left|sgn\big(s(n) - sgn(n-1)\big)\right| \qquad [7]$$

  Respecto a la energía, se calcula trama a trama, la energía máxima, mínima y media, de acuerdo con los parámetros establecidos en [19]. Posteriormente, tanto la energía como la tasa de cruces por cero se normalizan y se suman. Si la suma de ambos parámetros en la trama correspondiente es menor que cero, ese segmento se considera no sonoro. Finalmente, el parámetro se calcula como el número de segmentos no sonoros respecto al número total de segmentos de la señal.

<u>Parámetros de ruido:</u>

Se analizan en total seis parámetros relacionados con el ruido presente en la señal: relación ruido a harmónicos (NHR), relación harmónicos a ruido (HNR), energía normalizada de ruido(NNE), índice de turbulencia en la voz (VTI), índice de fonación (SPI) y ratio de la excitación glotal a ruido (GNE).

- NHR: mide la relación entre la energía de las componentes harmónicas en el rango 1500-4500Hz y la energía de las componentes harmónicas en el rango 70-4500Hz. Representa el ruido existente en la vocalización. [14]

- VTI: parámetro muy parecido al anterior. En esta ocasión mide la relación entre la energía de las componentes harmónicas de alta frecuencia respecto de la energía de los harmónicos en las áreas estables de la fonación. En este caso, consideramos el rango de altas frecuencias entre 2800 y 5800Hz y las áreas estables entre 70 y 4500Hz. Con este parámetro se consigue obtener las turbulencias de la señal presentes en las altas frecuencias. [14] [22]

- SPI: calcula la ratio entre la energía del harmónico de más baja frecuencia respecto de la energía del harmónico de más alta frecuencia. Da una idea del tipo de vocalización con la que se ha emitido la señal de voz. [14] [23]

- HNR: intenta medir de manera objetiva la sensación de ronquera en la voz y hay diversas interpretaciones de este parámetro. La primera de ellas es considerarlo la inversa del NHR. Otra aproximación es la relación harmónicos a ruidos en dominio Cepstral, en la ecuación 8. Esta aproximación calcula el parámetro para cada banda frecuencial, β, filtrando la señal con un filtro de banda eliminada, cuya longitud es la anchura del pico harmónico. La transformada de Fourier de la salida del filtro contiene la estimación de la energía de ruido presente en la señal. Finalmente, se calcula la diferencia entre la energía presente en los harmónicos y la energía de ruido. [4] [1]

$$\mathrm{HNR}_{\beta}(\mathrm{f}) = \mathrm{mean}\big(\mathrm{harmonic(f)}\big)_{\beta} - \mathrm{mean}\big(\mathrm{noise(f)}\big)_{\beta} \qquad [8]$$

  Aunque una de las primeras definiciones, establecida por Paul Boersma [24] se basa en la autocorrelación de la señal. Para señales estacionarias, el máximo global de la función de autocorrelación se encuentra en τ = 0. Si hay más máximos locales, se encontrarán en los múltiplos de T_0. Sabemos que el valor de la autocorrelación en cero corresponde con la potencia de la señal, entonces, si normalizamos la autocorrelación en τ = T_0, tendremos la potencia del componente harmónico. Su complementario será la potencia relativa de ruido, como sintetiza la fórmula 9, donde r, es la autocorrelación.

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

- GNE: la primera definición de este parámetro aparece en [25] para cuantificar la parte de señal producida por la excitación de las cuerdas vocales frente al ruido generador por turbulencias en la fonación. Para calcularlo, se baja la frecuencia de muestreo de la señal hasta 10kHz y han de tenerse en cuenta diferentes bandas frecuenciales con un ancho de banda de 1kHz y diferentes frecuencias centrales. El procesado se hace en tramas de 30ms con un solape de 10ms entre ellas.

  Tras bajar la frecuencia de muestreo, el siguiente paso es aplicar un filtrado inverso, usando un predictor lineal de orden 13. A continuación, se calculan las envolventes de Hilbert en cada una de las bandas frecuenciales y se

realiza la correlación cruzada entre los pares de las envolventes cuyas frecuencias centrales sean mayores que la mitad del ancho de banda. Luego se calcula el máximo entre los máximos de todas las correlaciones. Este máximo final es el parámetro GNE. [1]

- NNE: introducida por Kasuya et al. en 1986 [26] asume que las voces patológicas son más ruidosas que las voces normales. Para calcularlo, se baja la frecuencia de muestreo a 10kHz y se calcula el pitch en tramas de 40ms con un solape de 20ms. Con la mediana de los valores de pitch obtenidos se calcula un valor, M, que será el tamaño de la ventana de Hamming con la que se calculará cada trama. Se asume que cada trama enventanada, x_m está compuesta por una señal periódica, s_m y un ruido aditivo, w_m (Ecuación 10). La energía normalizada de ruido se calcula en el dominio frecuencial, siendo la expresión final la mostrada en la Ecuación 11, donde $\widehat{W}_m$ es la estimación del ruido y X_m es la transformada de Fourier de la parte periódica de la señal. N_h y N_l indican la banda frecuencial en la que se evalúa el parámetro y L es el número de tramas analizadas.

$$x_m(n) = s_m(n) + w_m(n) \qquad [10]$$

$$NNE = 10 * log\left[\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{m=1}^{L}\left|\widehat{W}_m(k)\right|^2 * \left(\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{m=1}^{L}|X_m(k)|^2\right)^{-1}\right] \qquad [11]$$

$\widehat{W}_m(k)$ es una estimación de la transformada de Fourier de la parte ruidosa de la señal. Teniendo en cuenta que $s_m(n)$ es periódica, $S_m(k)$ contribuye a la estructura periódica de $|X_m(k)|$. Al utilizar una ventana de Hamming cuyo tamaño es relativamente largo, la aportación de $|S_m(k)|$ se vuelve prácticamente inapreciable en los valles de la transformada de Fourier. De esta forma, el valor de $\widehat{W}_m(k)$ se calcula de forma diferente en las los valles y en los picos de la señal transformada. En los valles, directamente será el valor de $|X_m(k)|$ y en los picos se calculará con la fórmula 12, donde N es el número de puntos de la región del valle D_i.

$$\left|\widehat{W}_m(k)\right|^2 = \frac{1}{2}\left\{\sum_{r\in D_i}|X_m(r)|^2\,(N_i)^{-1} + \sum_{r\in D_i}|X_m(r)|^2(N_{i+1})^{-1}\right\} \qquad [12]$$

Los parámetros expuestos anteriormente tienen una doble función. La primera de ellas es analizar cada una de las nuevas grabaciones que se le pase al sistema para determinar si esa grabación contiene una patología o no. La segunda de ellas es crear un vector de entrada a la red neuronal. Los parámetros son almacenados en un

fichero en formato *csv* y a partir de la columna siete están guardados los parámetros. Las primeras columnas siguen la siguiente distribución:

Columna 1: '0' para voces sanas, '1' para voces patológicas.

Columnas 2 a 5: nombre de las patologías presentes, vacío en ausencia de patología. Se reservan cuatro columnas porque en la base de datos la grabación con más patologías devuelve cuatro afecciones.

Columna 6: tono de la vocal, '0' para alto, '2' para normal y '3' para bajo.

Columna 7: el sexo del sujeto, '0' para hombres y '1' para mujeres.

Se ha comentado que todos los parámetros devuelven un único valor numérico a excepción del MFCC que devuelve una matriz de MxN. Como este formato no se ajusta al formato de un fichero *csv*, los coeficientes del MFCC se ajustan a un vector de 1x(M*N). Aun así, todas las filas del fichero tienen que tener el mismo número de columnas, y como el tamaño de la matriz depende de la longitud de la señal a procesar, este parámetro se calcula con la señal recortada al número de muestras que posee el fichero de menor tamaño de la base de datos.

Cuando el fichero ya ha sido generado, se normaliza por columnas, de forma que los elementos tengan media nula y desviación típica igual a 1. Después, el valor medio de cada columna y su desviación típica se guardan en otro fichero, que se utilizará a la hora de proyectar el resultado final del proceso.

La última función que tienen estos parámetros es generar un gráfico orientativo en la herramienta final. El gráfico es de tipo radar, se muestra en escala logarítmica y muestra los valores normalizados y los valores normalizados medios de una persona sana.

2.3 Valoración de trastornos en el habla

La parte de valoración de trastornos del habla es la que evalúa si a partir de los parámetros descritos en el apartado anterior, la voz presenta una patología o no, con metodología de aprendizaje automático. Esta parte ha sufrido varias mutaciones desde su concepción final hasta la última versión programada.

La idea inicial era que la valoración se hiciese en dos etapas. La primera etapa decidiría la presencia o ausencia de patología y en caso de presencia, se pasaría a una segunda etapa que informaría que patologías eran las probables. Además, el proceso como ya se ha mencionado, sería separado para hombres y para mujeres.

La primera etapa sería un clasificador binario, y se probaron diversas opciones con la herramienta *Scikit-learn*. Entre estas opciones estaban *Support Vector Machines*, *Stochastic Gradient Descent Classification*, *Nearest Neighbours Classification* y *Gradient Boosting Classifier.* El fichero con el que se hicieron estas aproximaciones

contenía un 50% de voces patológicas y 50% de voces sanas, ambas de hombres, y la precisión estaba en torno al 74%.

La segunda etapa requería un clasificador multi-etiqueta, es decir, cada muestra del fichero no está restringida a una sola patología; es más, hay grabaciones con hasta cuatro patologías diagnosticadas y, por tanto, el clasificador debería detectar que patologías se ajustan a una misma grabación. Los resultados con clasificadores tradicionales (*SVM, Gradient Boosting Classifer*) fueron un tanto pobres, por lo que se hizo un reajuste del planteamiento inicial.

La nueva aproximación parte de hacer el proceso en una sola etapa, añadiendo a las voces sanas una etiqueta en la columna 2 del documento csv que indique '*healthy'*. De esta forma, con un solo clasificador multi-etiqueta el problema estaría resuelto, aunque en esta ocasión se recurre a una red neuronal artificial (ANN) con aprendizaje supervisado. Pero antes de proceder con la red, hay que transformar posibles salidas, ya que la red no comprende las etiquetas. La herramienta *MultiLabel Binarized* de *Scikit-Learn* genera una matriz binaria cuyo número de filas corresponde con el número grabaciones que se procesan y el número de columnas corresponde con el número total de patologías. Para cada muestra, marca con un '1' si la patología está presente.

La partición de la base de datos se compone de datos en entrenamiento y test. La parte de validación se realiza en el propio entrenamiento ya que la librería utilizada permite reservar un porcentaje de los datos para validar el funcionamiento de la red.

Una vez aclaradas como se organizan las particiones de entrenamiento y validación, el siguiente paso es calcular cuántas capas necesita la red neuronal artificial (ANN) y cuantas neuronas hay en cada capa. Según Jeff Heaton [27], en la mayoría de los problemas prácticos no hay razón para utilizar más de una capa oculta, y no hay un método exacto para calcular el número de neuronas, pero nos podemos guiar de la siguiente forma:

- El número de neuronas debería estar entre el número de parámetros de entrada y el número de salidas.
- El número de neuronas debería ser 2/3 del tamaño de la entrada, más el número de salidas.
- El número de neuronas debería ser menos que dos veces el número de entradas.

La selección final del número de neuronas se ha llevado a cabo mediante un barrido, entrenando ambas redes con un número de neuronas establecido siguiendo el criterio de Heaton. Se puede comprobar (Figura 2.5 y Figura 2.7) que la precisión de los datos de entrenamiento aumenta conforme aumenta en número de neuronas, pero que en validación el aumento de neuronas no conlleva a un mayor rendimiento. Por tanto, se ha elegido tener menos neuronas y mejorar el coste computacional.

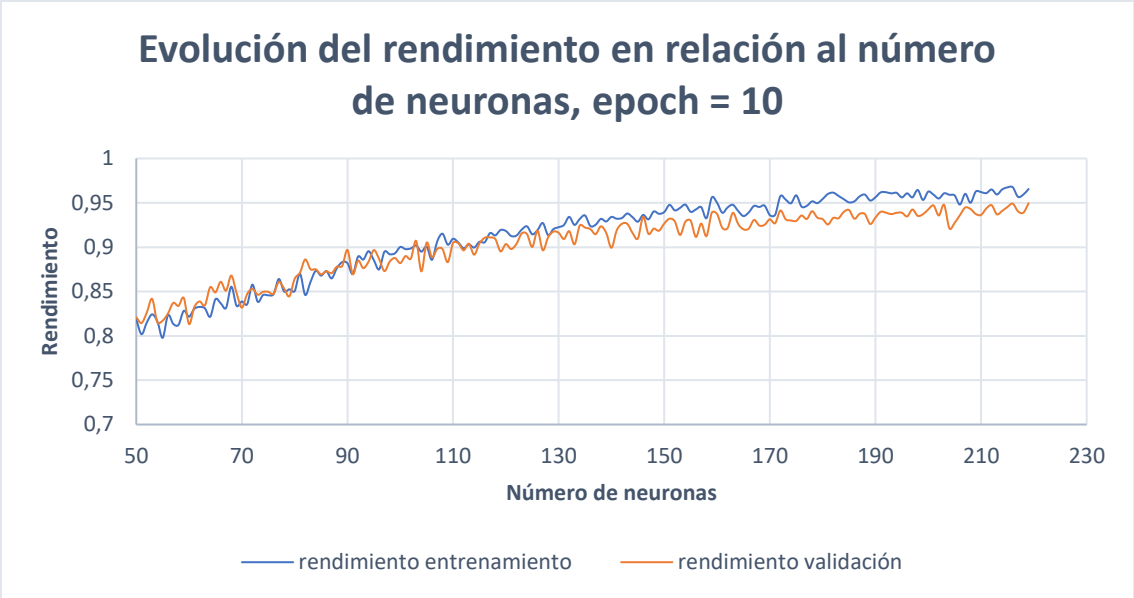

Figura 2.5 Evolución del rendimiento en función del número de neuronas. Hombres

Con todo lo anterior, esta vez utilizando el *framework* de Keras, se llega a una red neuronal con 168 neuronas en la capa oculta y la activación es de tipo tangente hiperbólica. La capa de salida contiene tantas neuronas como patologías se detecten en el fichero de entrada y la activación sigmoide permite que la red devuelva una lista con la probabilidad de cada patología en la muestra. Esta topología de red se muestra en la figura 2.6.

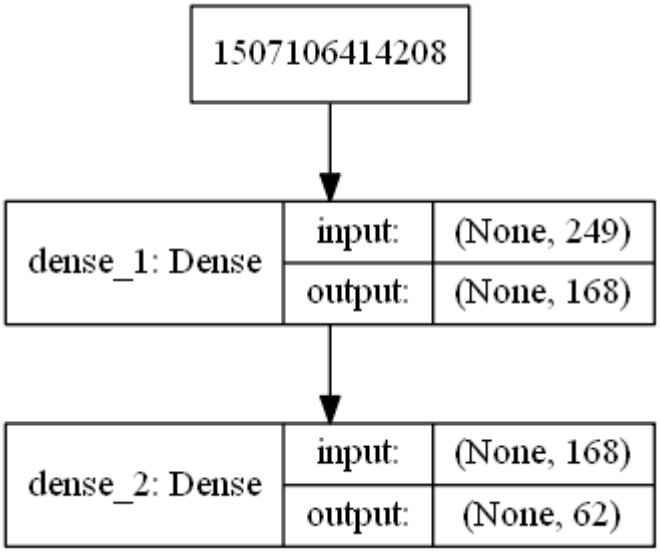

Figura 2.6 Topología de red

El mismo proceso se repite con la base de datos de voces femeninas. En este caso, la red neuronal cuenta con 161 neuronas en la capa oculta. Una vez obtenidas las especificaciones de ambas redes neuronales, se guarda el modelo y los generadores de la salida binarizada, para poder utilizarlo luego para predecir futuras muestras.

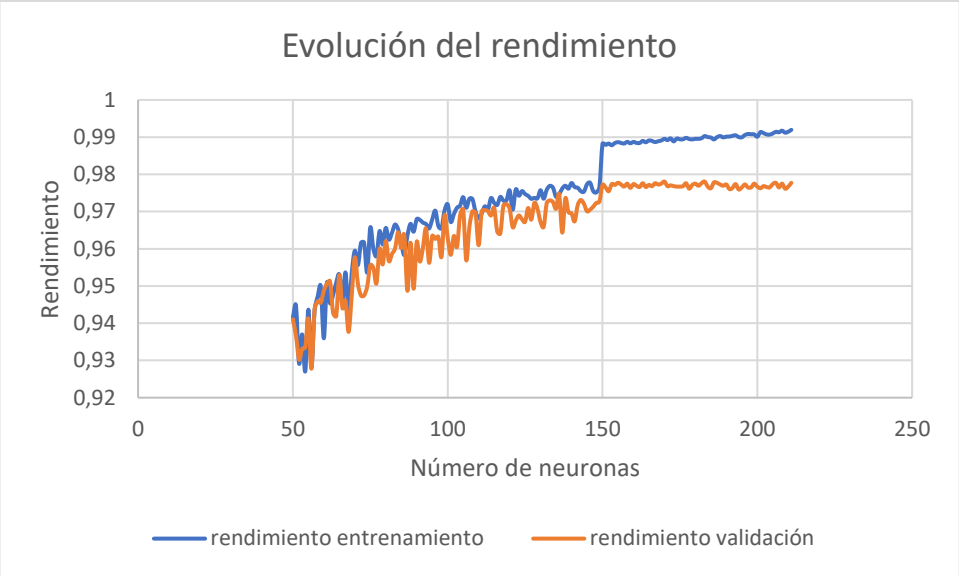

Figura 2.7 Evolución del rendimiento al incrementar el número de neuronas. Mujeres

Las métricas que se emplean para probar el rendimiento en el entrenamiento y en la validación de las redes neuronales son las proporcionadas por Keras: *accuracy* y *top_k_categorical_accuracy*. La primera de ellas compara cada elemento del vector de salida de la red neuronal con el vector de referencia. La segunda, compara el máximo del vector de referencia con las k probabilidades más altas devueltas por la red. Se han implementado dos métricas más *custom_1 y custom_2*, solo disponibles para la etapa de validación. Ambas ordenan las probabilidades de ocurrencia de las patologías de mayor a menor. Como en la etapa de validación sabemos el número de patologías que tiene la muestra, la primera aproximación busca que la patología deseada sea la más probable. La segunda métrica se "conforma" con que la patología deseada esté entre las cuatro más probables.

Por último, la mayor parte de los recursos empleados para el desarrollo de este apartado provienen de la propia documentación de *Keras* [9] y *Scikit-Learn* [8] y de otros recursos web como son [28] y [29]

2.4 Prueba de concepto final

Como colofón a las secciones anteriores, se ha creado una herramienta que permite enlazar el análisis automático de la señal de voz con la red neuronal artificial de forma tangible. El ciclo de vida de esta herramienta es el que aparece en la Figura 2.8. Para cada nueva muestra de voz, se le realiza el análisis automático y tras pasarla por la red neuronal ya entrenada, ésta debería ser capaz de devolver una lista con las cuatro patologías más probables que presenta la red neuronal como se muestra en la Figura 2.9.

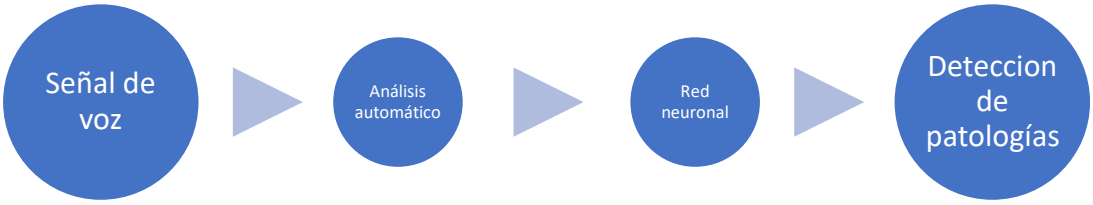

Figura 2.8 Ciclo de vida de la aplicación

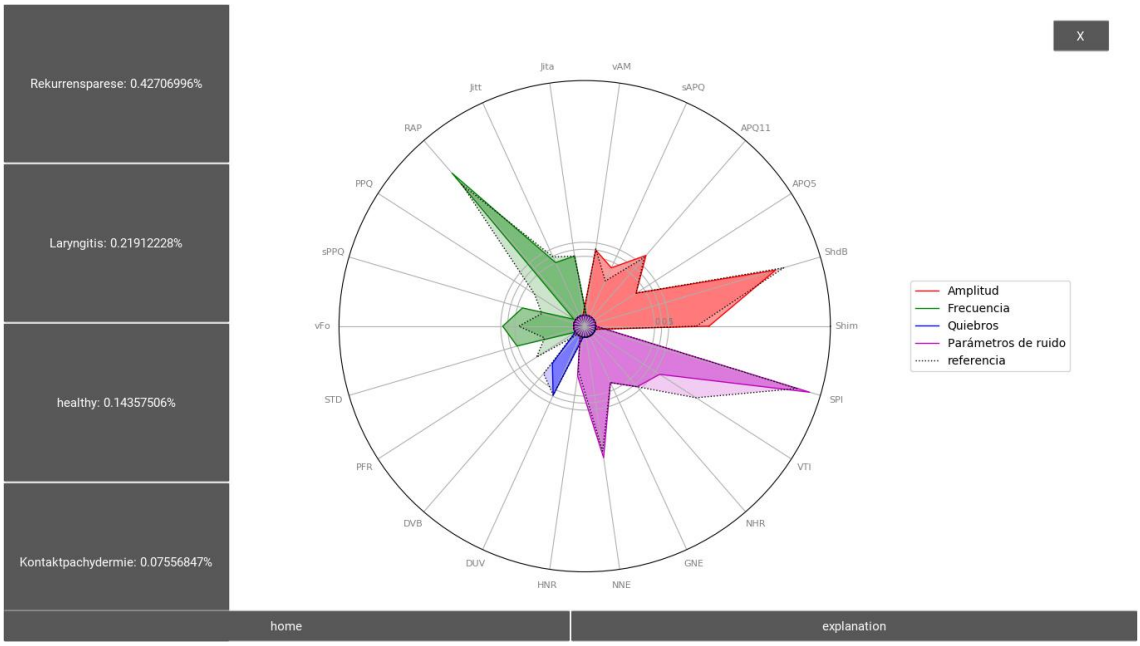

Figura 2.9 Ejemplo de aplicación

La herramienta es muy sencilla e intuitiva. En primer lugar, hay que indicar si la muestra de voz es femenina o masculina con el menú de la Figura 2.10 y después, permite elegir el fichero de audio deseado. Tras procesar el fichero, devuelve un gráfico con el valor de cada parámetro y las patologías más probables.

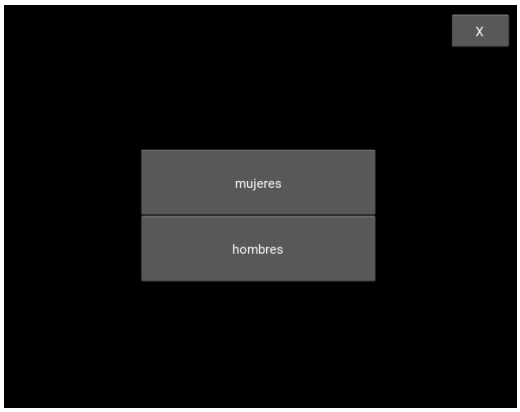

Figura 2.10Menú de selección

La herramienta, al ser una prueba de concepto se ha desarrollado como una aplicación de escritorio para facilitar su testeo.

3. Resultados

Los resultados de este proyecto los podemos separar en dos partes. Por un lado, analizaremos los resultados de la parte de valoración de trastornos del habla y por otro, los relacionados con la aplicación.

Los resultados obtenidos en la valoración de trastornos en el habla dan lugar a diversas interpretaciones, ya que dependen en gran medida de las métricas utilizadas y no es fácil ajustar una métrica para un clasificador multi-etiqueta, con más de 40 etiquetas en ambas redes. Debido a esta circunstancia, se van a analizar los resultados obtenidos en los cuatro espacios muestrales que se han procesado.

Como paso previo, recordar que solo se tiene en cuenta la letra /a/ en sus diferentes tonos. Los candidatos para formar esta base de datos de entrenamiento se han seleccionado de forma aleatoria, pero asegurando que cada patología esté al menos una vez, sin importar el tono de la grabación. En total, para cada grabación se extraen 21 parámetros, que componen las 249 características de entrada de la red neuronal.

a) Base de datos de hombres, con equilibrio entre voces sanas y patológicas.
   En la tabla 3.1 se muestran los resultados medios obtenidos con esta base de datos y un entrenamiento de 168 neuronas. La primera métrica devuelve valores sorprendentemente altos, ya que valora positivamente que las patologías no presentes en la muestra original tengan una probabilidad muy baja en el vector de predicciones. Por tanto, no es una métrica muy fiable para la detección de patologías, pero sí que podría ser una buena métrica para descartar las patologías menos probables.

Métrica	*Entrenamiento (%)*	*Validación (%)*
Accuracy	99,01	97,82
top_k_categorical_accuracy	98,59	72,69
Custom_1	-	54,36
Custom_2	-	73,24

Tabla 3.1: Resultados de la aproximación a)

Los datos de las dos primeras filas de la tabla se pueden ver con mayor detalle en la figura 3.1. La precisión de la primera métrica crece muy rápido con un número de *epoch* o vueltas muy pequeño, y se mantiene en valores en torno al 99%. La segunda métrica, crece más despacio y en la validación se mantiene estable en torno a valores del 70%. Por otra parte, la métrica Custom_1 indica si la patología deseada es la más probable entre los valores que devuelve la evaluación. La métrica Custom_2, indica la tasa de éxito con la que la patología deseada está entre las cuatro más probables. Como la aplicación devuelve las cuatro patologías más probables para ofrecer una guía del estado de la voz, esta puede ser la métrica más adecuada.

El modelo de pérdidas de la figura 3.2, muestra como las pérdidas decrecen hasta estabilizarse conforme la red va entrenando.

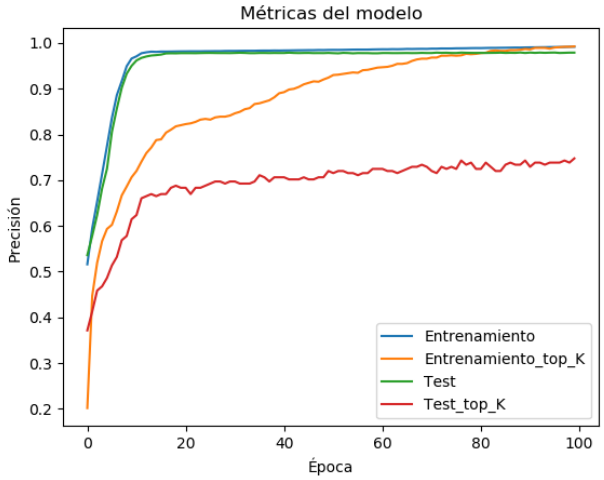

Figura 3.1 Métricas del modelo a)

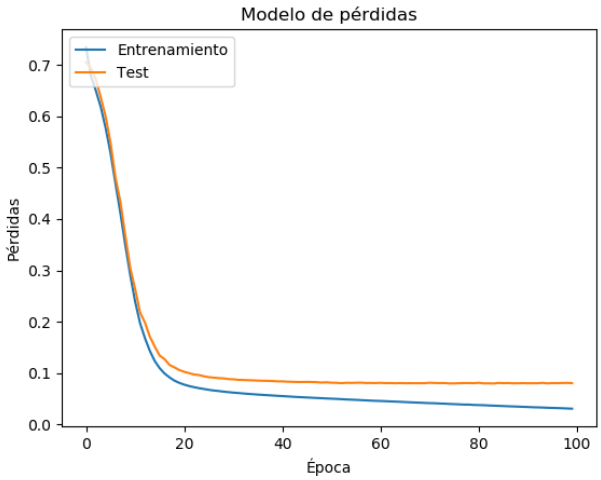

Figura 3.2: Modelo de pérdidas del modelo a)

b) Base de datos de hombres, utilizando el 80% de las voces patológicas y la totalidad de las voces sanas, entrenada con 168 neuronas.

Métrica	*Entrenamiento (%)*	*Validación (%)*
Accuracy	98,94	98,28
top_k_categorical_accuracy	95,35	69,66
Custom_1	-	44,10
Custom_2	-	69,04

Tabla 3.2Modelo de precisión para la aproximación b)

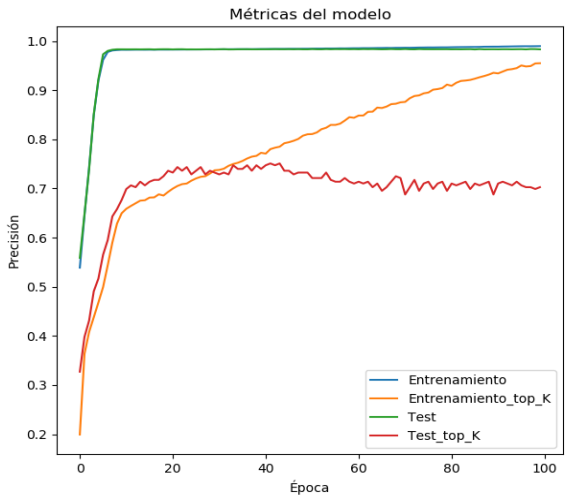

Figura 3.3: Métricas del modelo b)

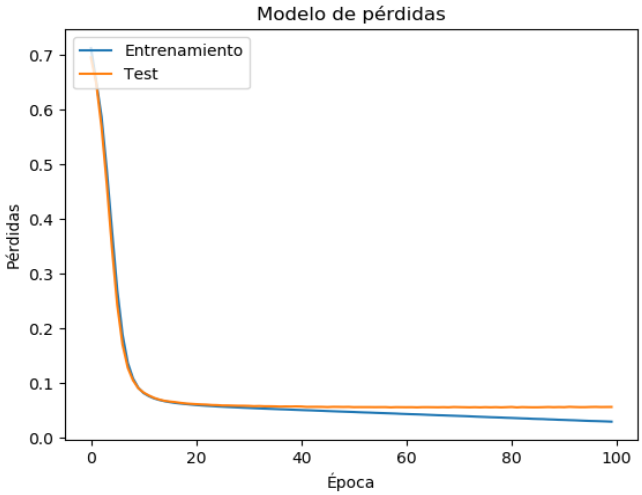

Figura 3.4: Modelo de pérdidas del modelo b)

Se observa en la Tabla 3.2 un rendimiento ligeramente inferior al anterior. El incremento de la base de datos no ha supuesto una mejora sustancial. Esto puede deberse a que hay patologías infrarrepresentadas en la base de datos original, por tanto, aunque la muestra extraída ahora sea más amplia, no

consigue ofrecer más muestras de esas patologías. En las Figuras 3.3 y 3.4 se muestra el entrenamiento de la red neuronal y la evolución de las pérdidas conforme avanza el entrenamiento.

c) Base de datos de mujeres, con equilibrio entre voces sanas y patológicas.
Se repiten los resultados obtenidos en la base de datos de hombres en la tabla 3.3. La primera métrica es muy optimista, al considerar las probabilidades de todas las patologías y no solamente la de la patología presente en la muestra. La métrica *custom_1* es la más pesimista de nuevo, con aproximadamente un 51% de aciertos. Las otras dos métricas ofrecen resultados muy parecidos e indican que la patología presente en la señal de voz se encuentra entre las cuatro con mayor probabilidad. Estos resultados se obtienen con 161 neuronas.

Métrica	*Entrenamiento (%)*	*Validación (%)*
Accuracy	98,89	97,76
top_k_categorical_accuracy	97,51	76,29
Custom_1	-	51,34
Custom_2	-	72,57

Tabla 3.3Modelo de precisión para la aproximación c)

En las figuras 3.5 y 3.6 se muestran el entrenamiento de la red y el modelo de pérdidas. Ambos repiten el patrón de las voces masculinas, bajando las pérdidas conforme el modelo va entrenando, mientras que la precisión se estabiliza.

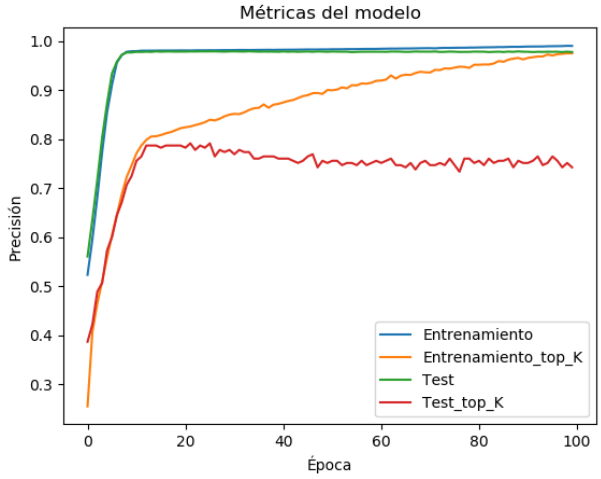

Figura 3.5: Métricas del modelo c)

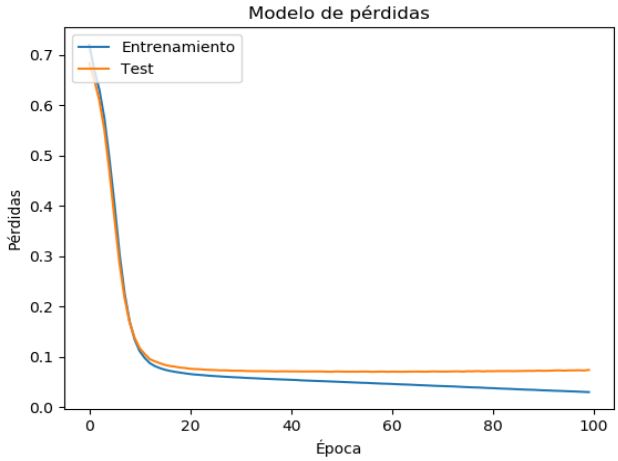

Figura 3.6: Modelo de pérdidas del modelo c)

d) Base de datos de mujeres, utilizando el 80% de las voces patológicas y la totalidad de las voces sanas.
Al igual que ocurre al aumentar la representación de las voces masculinas en el apartado b), un aumento de la representación de patologías no mejora el rendimiento de la red neuronal. Como se ha indicado, esto puede deberse a la infrarrepresentación de algunas patologías en la base de datos inicial. La red se ha entrenado con 194 neuronas en esta ocasión.

Métrica	*Entrenamiento (%)*	*Validación (%)*
Accuracy	99,08	97,55
top_k_categorical_accuracy	98,94	74,26
Custom_1	-	45,12
Custom_2	-	72,07

Tabla 3.4 Modelo de precisión para la aproximación d)

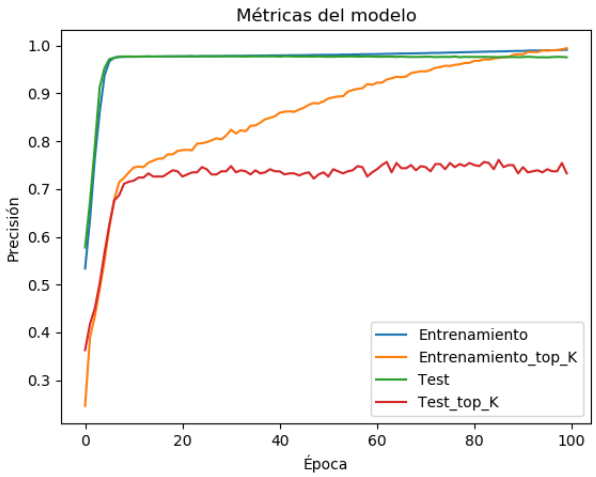

Figura 3.7: Métricas del modelo d)

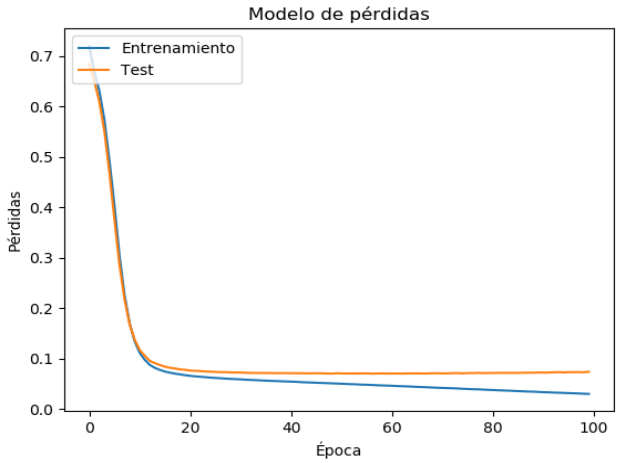

Figura 3.8: Modelo de pérdidas del modelo d)

Además de la infrarrepresentación de algunas patologías en la base de datos, estos resultados pueden deberse a otro fenómeno. Para las voces femeninas, contando los tres tonos de la vocal /a/, hay un total de 1245 grabaciones de voces sanas y 2159 de voces patológicas. El clasificador está considerando la voz sana como una patología más que clasificar, y aunque se equilibre la muestra con el mismo número de voces patológicas que sanas, la patología 'voz sana' es la más común, por tanto, puede haber un sesgo hacia este resultado.

Como estos resultados tienen margen de mejora, se ha propuesto probar la red neuronal con una muestra más selecta. Se han seleccionado las patologías que tuviesen más de 19 sesiones grabadas y las voces sanas se han limitado al mismo número de grabaciones que la patología más frecuente, a saber, *Rekurrensparese para hombres* y Disfonía hiperfuncional en mujeres.

Con esta muestra más reducida, se han conseguido subir ligeramente los rendimientos en validación tanto en hombres como para mujeres. Las métricas custom_2 y top_k_categorical_accuracy están en torno a 76-77% tanto en voces masculinas como en femeninas, aunque se han conseguido picos del 82%. Para las voces femeninas se han obtenido el máximo rendimiento con una red neuronal con 106 neuronas en la capa oculta y sus métricas se muestran en la tabla 3.5. En la tabla 3.6 se muestran los resultados de las voces masculinas obtenidos con una red de 218 neuronas.

Métrica	*Entrenamiento (%)*	*Validación (%)*
Accuracy	96,38	91,89
top_k_categorical_accuracy	97,92	77,21
Custom_1	-	45,96
Custom_2	-	77,68

Tabla 3.5 Métricas medias del grupo reducido de voces femeninas

Métrica	*Entrenamiento (%)*	*Validación (%)*
Accuracy	97,78	94,08
top_k_categorical_accuracy	99,01	75,64
Custom_1	-	47,52
Custom_2	-	75,94

Tabla 3.6 Métricas medias del grupo reducido de voces masculinas

Respecto a la aplicación, finalmente solo se ha creado la versión de escritorio, debido a incompatibilidades entre las librerías utilizadas en el proyecto y *buildozer*, que es la herramienta que genera la aplicación en Android.

4. Conclusiones

Las técnicas de aprendizaje automático unidas al procesado automático de señales ofrecen un amplio abanico de posibilidades en el tratamiento de señales de voz. En este trabajo se ha realizado una prueba de concepto para estudiar la capacidad de estas técnicas para la valoración de trastornos en el habla.

Respecto a la parte de análisis automático de la señal de voz, se han analizado parámetros acústicos que ofrecen cierta relevancia sobre el estado de la voz. Aunque estos ofrecen buenas prestaciones, se podría estudiar la inclusión de más parámetros que permitan medir otras características de la señal, como puede ser el temblor, para tener un análisis más completo de la voz a estudiar.

La parte de valoración de trastornos del habla ha sufrido muchas mutaciones desde la idea original y se han probado multitud de clasificadores hasta optar por el modelo de red neuronal artificial. En las primeras aproximaciones, se realizaba una primera clasificación binaria que decidía si una voz era patológica o no. Un segundo clasificador decidía qué patología podía estar presente. El cambio de paradigma se dio porque la primera red neuronal ofrecía un rendimiento en torno al 70%, es decir, que la probabilidad de la segunda red estaría condicionada al resultado de la primera. Por este motivo se decidió el cambio a una topología de una única red neuronal, para evitar la probabilidad condicionada de los dos eventos.

Por otra parte, la unión del análisis automático de la señal de voz y la valoración de trastornos del habla han dado lugar a una herramienta en forma de aplicación de escritorio que permite testear rápidamente nuevas grabaciones e indicar que patología o patologías son más probables o en su defecto si el sujeto de la grabación está sano. Aunque la detección de patologías ofrece una precisión entorno al 72-74%, la herramienta base ya está creada y se podrían mejorar estos rendimientos de varias maneras. La primera de ellas sería ampliando la base de datos, sobre todo para las patologías menos frecuentes. La segunda, restringiendo más la base de datos para que la herramienta reconozca menos patologías, pero con mayor precisión. Por otra parte, también se podría mejorar la red neuronal, incorporando redes neuronales convolucionales o con más neuronas en las capas ocultas.

Por último, hay que destacar el coste económico del proyecto. Esta prueba de concepto no ha requerido de la utilización de software propietario ya que todos los parámetros a excepción de los MFCC se han programado en Python y la base de datos es de acceso libre y gratuito. Es decir, se ha conseguido un prototipo de una herramienta sin costes en licencias.

5. Líneas futuras

El principal elemento que ha quedado en el aire en este proyecto ha sido trasladar la aplicación a entornos móviles con sistema operativo Android. De este modo, su uso no se limitaría a procesar voces pregrabadas, sino que se abriría al procesado en tiempo real. Por tanto, en una posible revisión de este proyecto esta sería la primera línea que seguir.

Por otra parte, una ampliación de la base de datos permitiría detectar más patologías, haciendo hincapié en ampliar la base de datos con voces de niños y niñas, ya que en este estudio no había una muestra representativa de ellos y, por tanto, no se han analizado.

Otra posible línea futura sería aplicar técnicas de descomposición, como PCA (*Principal Component Analysis*), para reducir el número de características de entrada a la red neuronal y así mejorar los costos computacionales al tener una red más pequeña y fácil de procesar.

6. Bibliografía

[1] David Martinez, Eduardo Lleida, Antonio Miguel, Alfonso Ortega, «Automatic GRBAS Rating for Voice Quality Assessment Using Multidimensional Information,» *Speech Communication*, 2014.

[2] D. Martínez, E. Lleida, A. Ortega, A. Miguel y J. Villaba, «Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit,» de *Proceedings of Advances in Speech and Language Technologies for Iberian Languages*, Madrid, Iberspeech, pp. 99-109.

[3] A. Tsanas y P. Gomez-Vilda, «Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels,» de *Multidisciplinary Conference of Users of Voice, Speech and Singing*, Las Palmas de Gran Canaria, 2013.

[4] R. B. Reilly, R. Moran y P. Lacy, «Voice Pathology Assesment based on a Dialogue System and Speech Analysis,» de *AAAI Fall Symposium*, 2004.

[5] P. Harar, J. B. Alonso-Hernández, J. Mekyska, Z. Galaz, R. Burget y Z. Smekal, «Voice Pathology Detection Using Deep Learning: a Preliminary Study,» de *International Conference and Workshop on Bioinspired Intelligence*, Funchal, 2017.

[6] «https://kivy.org,» [En línea]. [Último acceso: 18 Noviembre 2018].

[7] «https://www.tensorflow.org/?hl=es,» [En línea].

[8] «https://scikit-learn.org/stable/index.html,» [En línea].

[9] «https://keras.io,» [En línea].

[10] «https://conda.io/docs/,» [En línea].

[11] «https://www.jetbrains.com/pycharm/,» [En línea].

[12] «Saarbruecken Voice Database,» [En línea]. Available: http://stimmdb.coli.uni-saarland.de.

[13] A. Tsanas, M. A. Little, C. Fox y L. O. Raming, «Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, nº 1, pp. 181-190, 2014.

[14] *Disorderd Voice Database Model 4337, Operations Manual*, Boston: Lab, Massachusetts Eye and Ear Infirmary Voice and Speech, 2008.

[15] «Python speech features,» [En línea]. Available: https://python-speech-features.readthedocs.io/en/latest/#.

[16] «Mel Frequency Cepstral Coefficient (MFCC) tutorial,» [En línea]. Available: http://www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

[17] D. W. Paul Boersma, «Praat Manual,» [En línea]. Available: http://www.fon.hum.uva.nl/praat/manual/Voice_1__Voice_breaks.html.

[18] B. S. Atal y L. R. Rabiner, «A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition,» *EEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, nº 3, pp. 201-212, 1976.

[19] A. M. Kondoz, Digital Speech. Coding for Low Bit Rate Communication Systems, John Wiley & Sons, Ltd, 2004.

[20] M. Hodgkinson, «CS425 Audio and Speech Processing,» 2012.

[21] D. O'Shaughnessy, Speech communications: human and machine, New York: IEEE Press, 2000.

[22] B. J. Poburka y C. Buchkoski, «Voice Turbulence Index: Normative Data,» [En línea].

[23] B. J. Poburka y M. Riesgaard, «Soft Ponation Index: Normative Data,» [En línea].

[24] P. Boersma, «Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled sound,» *Proc. Institute of Phonetic Sciences University of Amsterdam,* vol. 17, pp. 97-110, 1993.

[25] D. Michaeis, T. Gramss y H. Strube, «Glottal-to-Noise Excitation Ratio,» *Acustica/Acta acustica,* vol. 83, pp. 700-706, 1997.

[26] H. Kasuya, S. Ogawa, K. Mashima y S. Ebihara, «Normalized Noise Energy as an Acoustic Measure to Evaluate Pathologic voice,» *Acoustic Society of America,* vol. 80, nº 5, 1986.

[27] J. Heaton, «Heaton Research,» 2008. [En línea]. Available: https://web.archive.org/web/20140721050413/http://www.heatonresearch.com/node/707.

[28] J. Brownlee, «Machine Learning Mistery,» [En línea]. Available: https://machinelearningmastery.com.

[29] A. Rosebrock, «py Image Search,» [En línea]. Available: https://www.pyimagesearch.com.

Anexo A. Relación de patologías

En la tabla siguiente se muestran todas las patologías que componen la base de datos SVD. Se incluye su nombre original y su traducción en castellano, en los casos en los que se ha encontrado dicha traducción.

Nombre Patología	**Traducción**
Amyotrophe Lateralsklerose	Esclerosis lateral amiotrófica
Aryluxation	Aryluxation
Balbuties	Tartamudeo
Bulbarparalyse	Parálisis bulbar
Carcinoma in situ	Carcinoma in situ
Chondrom	Condroma
Chordektomie	Cordectomía
Cyste	Quiste
Diplophonie	Diplofonía
Dish-Syndrom	Hiperostosis esquelética idiopática difusa
Dysarthrophonie	Disartrofonía
Dysodie	Disodia
Dysphonie	Disfonía
Dysplastische_Dysphonie	Disfonía displásica
Epiglottiskarzinom	Carcinoma de epiglotis
Fibrom	Fibroma
Frontolaterale Teilresektion	Resección parcial frontolateral
Funktionelle Dysphonie	Disfonía funcional
GERD	Enfermedad por reflujo gastroesofágico
Gesangsstimme	Voz de canto
Granulom	Granuloma
Hyperasthenie	Hiperestesia
Hyperfunktionelle Dysphonie	Disfonía hiperfuncional
Hypofunktionelle Dysphonie	Disfonía hipofuncional
Hypopharynxtumor	Tumor de hipofaringe
Hypotone Dysphonie	Disfonía hipotónica
Internusschwache	Debilidad Interna
Intubationsgranulom	Granuloma de intubación
Intubationsschaden	Daño de intubación
Juvenile Dysphonie	Disfonía Juvenil
Kehlkopftumor	Tumor de laringe
Kontaktpachydermie	Kontaktpachydermie
Laryngitis	Laringitis
Laryngozele	Laringocele
Leukoplakie	Leucoplasia
Mediale Halscyste	Quiste cervical mediano
Mesopharynxtumor	Tumor de nasofaringe
Monochorditis	Monocorditis vasomotora

Morbus Down	Enfermedad Down
Morbus Parkinson	Enfermedad de Parkinson
Mutatio	Cambio
Mutationsfistelstimme	Cambio de falsete
N laryngeus superior Lasion	Lesión a los nervios laríngeos
N laryngeus superior Neuralgie	Neuralgia del nervio laríngeo superior
Non-fluency-Syndrom	El síndrome de la no fluidez
Orofaciale Dyspraxie	Dispraxia orofacial
Papillom	Papiloma
Phonasthenie	Fonastenia
Phonationsknotchen	Fonación sknötchen
Poltersyndrom	Síndrome espasmofénico
Psychogene Dysphonie	Disfonía psicógena
Psychogene Mikrophonie	Microfonía psicógena
Reinke Odem	Edema de reinke (edema en las cuerdas vocales)
Rekurrensparese	Parálisis recurrente- daño al nervio laríngeo
Rhinophonie aperta	Rinofonía abierta
Rhinophonie clausa	Rinofonía cerrada
Rhinophonie mixta	Rinofonía mixta
Sangerstimme	Sangerstimme
Sigmatismus	Sigmatismo interdental
Spasmodische_Dysphonie	Disfonía espasmódica
Stimmlippenkarzinom	Carcinoma de las cuerdas vocales
Stimmlippenpolyp	Pólipo en los pliegues vocales
Synechie	Sinequia vulvar
Taschenfaltenhyperplasie	Hiperplasia de pliegues vestibulares
Taschenfaltenstimme	Taschenfaltenstimme
Valleculacyste	Quiste de vallécula
Velopharyngoplastik	Velo faringoplastia
Vox senilis	Voz senil
Zentral-laryngale Bewegungsstorung	Trastorno del movimiento laríngeo central
Dysplastischer Kehlkopf	Laringe displásica