Ciclos respiratorios en audios de Auscultación

Santiago Ortiz Ceballos[1], Santiago Cardona Flórez[1].
Universidad de Antioquia[1].

***Abstract*— This document contains the analysis of a Kaggle database on auscultation of the lungs. The available audios are cleaned with digital linear filters, and non-linear ones like the Wavelet filter, some indices are calculated for their characterization and an Excel table is generated with the collected information, which is statistically analyzed, finding as a result that the healthy respiratory cycles differ from the sick ones in the average spectral power of their signal.**

Keywords: Pulmonary auscultation, Digital filters, Wavelet filter, Spectral analysis, Descriptive statistics

***Resumen*— Este documento contiene el análisis de una base de datos de Kaggle sobre auscultación pulmonar. Los audios disponibles se limpian con filtros digitales lineales, y no lineales como el filtro de Wavelet, se calculan algunos índices para su caracterización y se genera una tabla de Excel con la información recopilada, que es analizada estadísticamente, encontrando como resultado que los ciclos respiratorios sanos difieren de los enfermos en el promedio de la potencia espectral de su señal.**

Palabras clave: Auscultación pulmonar, Filtros digitales, Filtro Wavelet, Análisis Espectral, Estadística descriptiva.

I. INTRODUCCIÓN

La Auscultación es una técnica que permite escuchar los sonidos pulmonares a través del estetoscopio. El análisis de estos sonidos permite identificar rasgos característicos de las diferentes patologías que afectan los pulmones y/o las vías respiratorias en humanos. Sin embargo, escuchar estos sonidos no siempre es suficiente para brindar un diagnóstico acertado, por lo que se requieren estudios computacionales de las señales.

Actualmente se busca generar herramientas de análisis con el objetivo de agilizar el diagnóstico. Por ejemplo, en la Universidad de Tromso se realizó una tesis para la automatización de este análisis [1] y se tomó como base para el desarrollo de este estudio. El cual consiste en el procesamiento de señales de audios de Auscultación tomados de una base de datos de *Kaggle* sobre sonidos respiratorios [2]. El proceso incluye varias etapas de filtrado, un análisis en el dominio del tiempo y de la frecuencia, el cálculo de algunos índices de interés mencionados en la tesis para la caracterización de los ciclos respiratorios, la construcción de una tabla de datos que pueda ser analizada estadísticamente, un análisis exploratorio de datos, y finalmente la aplicación de técnicas de estadística descriptiva y pruebas de hipótesis.

En concreto, este estudio pretende identificar diferencias entre ciclos respiratorios sanos y enfermos, con miras al diseño de un equipo de clasificación y predicción para las señales de Auscultación.

II. MATERIALES Y MÉTODOS

Los materiales empleados en el desarrollo de este estudio fueron la base de datos de sonidos de Auscultación, que contiene información sobre 126 pacientes, donde se incluye información básica sobre cada uno de ellos; así como su diagnóstico y formato de adquisición de señales, las cuales fueron registradas a **4000** Hz y cada una de ellas incluye un archivo de anotaciones que indica inicio y fin de los ciclos respiratorios presentes en cada audio, y sí dicho ciclo presenta **Crepitancias y/o Sibilancias**. También se emplea la Tesis previamente mencionada, que se tomó como referencia, y el Paquete de Anaconda-Python para el procesamiento y análisis de los datos.

Inicialmente se realizó una **rutina para una señal de Auscultación**, la cual parte de aplicar un Filtro Pasa Banda con frecuencias de corte en **100** y **1000** Hz que es el rango de mayor interés en las señales de sonidos pulmonares.[2] Luego se aplica un filtro **Wavelet** con nivel 4 de descomposición, tomando el umbral Minimax y aplicando una ponderación Multinivel con filtro de forma Suave, para remover los sonidos cardiacos de la señal pulmonar. Sin embargo la señal resultante del Wavelet realmente elimina los sonidos pulmonares y deja los cardíacos, por lo que se toma la **diferencia** de la señal original con esta, para obtener como resultado una señal únicamente con sonidos pulmonares.

Posteriormente, se aplicó el filtro pasabanda aplicado al inicio, dado que hacer la diferencia entre las señales original, y la del corazón, realza nuevamente frecuencias bajas y altas que no son de interés.

Cabe destacar que los parámetros de filtro se establecieron con base en pruebas de ensayo y error, analizando los espectros de las señales pre/post filtrado y verificando los audios resultantes exportándolos al formato .wav para ser escuchados.

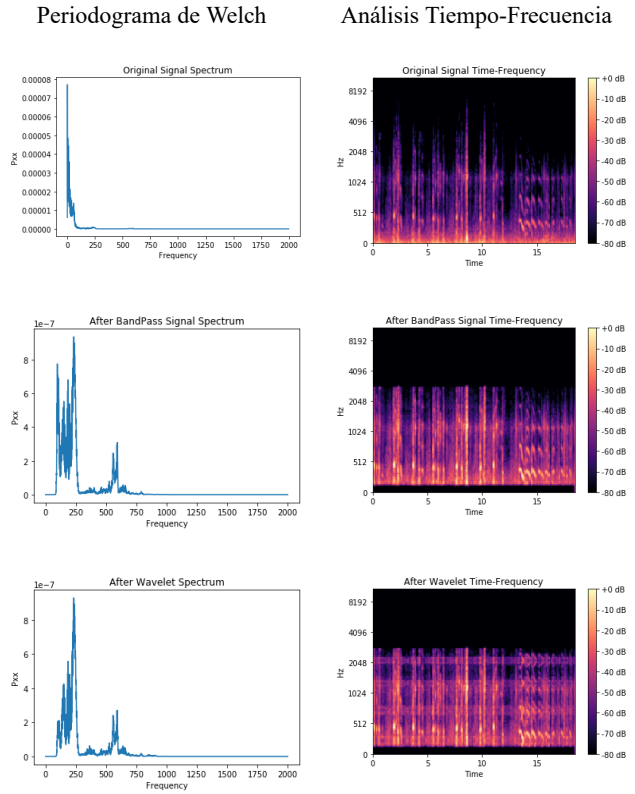
Una vez, se obtiene la señal “limpia” se procede con el cálculo de los índices de interés mencionados en la tesis, los cuales son la **Varianza**, el **Rango**, la suma del **promedio móvil** simple (que se divide en 2: un promedio Grueso y uno Fino) y el **promedio del espectro** de la señal.

Luego de definir la rutina para procesar una señal, se procede con el diseño de una rutina que procese la base de datos completa a medida que va generando una tabla que contenga la información del paciente, su diagnóstico, el inicio y fin de cada ciclo respiratorio, y los índices correspondientes a cada uno. Esto con el fin de exportar la información completa a un archivo de Excel para su análisis.

Finalmente, la tabla de Excel generada se importa desde un Jupyter Notebook en el cual se aplican técnicas de análisis exploratorio de datos, estadística descriptiva y prueba de hipótesis, para estudiar 2 casos: en el primero se plantea como hipótesis nula que los ciclos respiratorios **sanos**, presentan un **promedio móvil fino** similar al de un ciclo que tiene **Crepitaciones y Sibilancias**, y como hipótesis alternativa se plantea que estos ciclos presentan un promedio móvil fino diferente. En el segundo, se plantea como hipótesis nula que los ciclos respiratorios **sanos** presentan un **promedio del espectro** similar a los ciclos con **Crepitaciones y Sibilancias**; y como hipótesis alternativa, esta característica es diferente para ambos ciclos.

III. RESULTADOS

En las siguientes 6 figuras se presenta el procesamiento realizado paso a paso de una señal de audio.



Figuras 1:6. *Las filas 1, 2 y 3 presentan la señal original, después del pasabanda, y después del Wavelet respectivamente. En la columna izquierda se presentan los periodogramas de Welch y en la derecha el Wavelet Continuo del análisis tiempo-frecuencia.*

En la figura 7 se comparan las señales de audio resultantes de cada proceso.

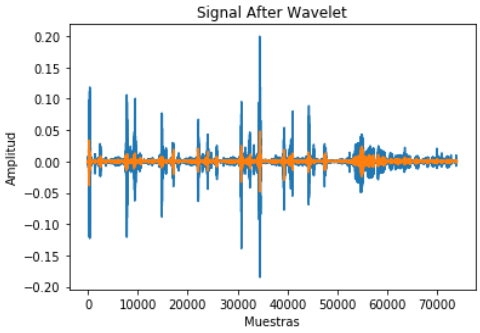
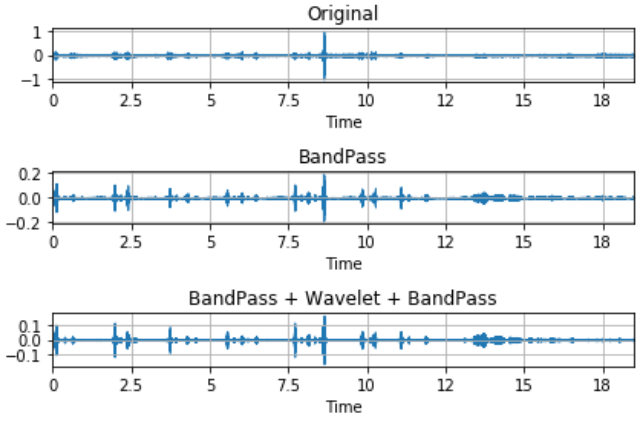


Figura 7. *Señal de audio antes (color azul) y después (color naranja) de aplicar el Wavelet.*



Figuras 8. *Comparación entre las señales resultantes.*

En el anexo llamado **Statistics** se presenta un análisis detallado de los datos, por lo que en está sección simplemente se tomarán los aspectos más relevantes que permiten desarrollar la prueba de hipótesis para los casos planteados.

En las figuras 9 y 10 se presenta como se encuentra distribuido el promedio de la sumatoria móvil fina y el promedio de la potencia espectral en 4 clasificaciones realizadas a partir de los datos recopilados que se detallan en el anexo **Statistics**.

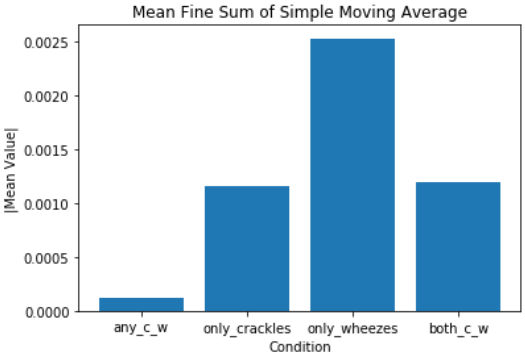


Figura 9. *Promedio de la Sumatoria Móvil Fina para cada categoría.*

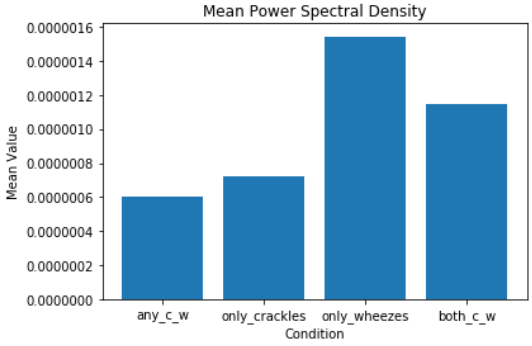
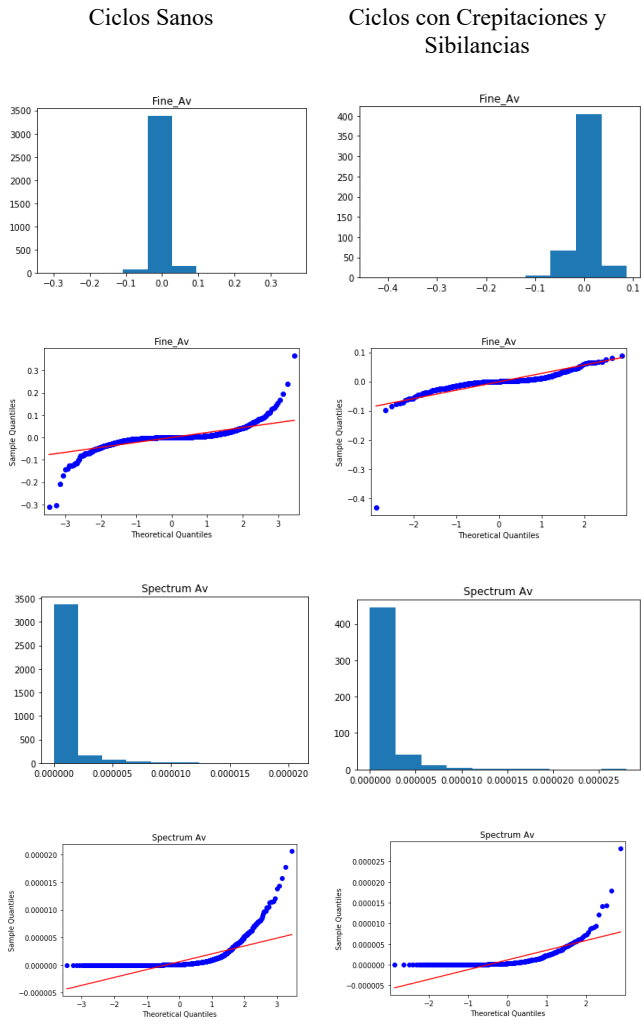


Figura 10. *Promedio de la Potencial Espectral para cada categoría.*

En las figuras 11 a 18 se presentan los histogramas de los índices bajo análisis para los dos casos planteados, con sus respectivos gráficos Cuartil-Cuartil que permiten corroborar la distribución de los datos.



Figuras 11:18. *Las filas 1 y 3 presentan los histogramas, mientras que las filas 2 y 4 presentan la relación cuartil-cuartil.*

Los resultados de los test aplicados para cada caso, se presentan tabulados en la tabla 1.

Tabla 1. *Resultados para las pruebas de hipótesis.*

Test No Param.	Spearman		Mannwhitneyu	
	Correlación	Valor p	Estadístico	Valor p
Caso 1	5.18%	28.40 %	90933	33.85 %
Caso 2	9.94%	3.9 %	62830	2.11 x 10^{-14} %

IV. Discusión

Los espectros resultantes del filtrado de las señales evidencian las frecuencias que son de interés en los sonidos pulmonares que se encuentran concentradas entre 100 y 2000 Hz como indica la literatura, siendo las de mayor energía aquellas entre 100 y 1000 Hz como muestran las paletas de colores y las cuales son de interés en este tipo de estudio.

Respecto a las distribuciones de los datos, el promedio móvil Fino tiende a caracterizarse como normal, sin embargo, como la población de sanos presenta muchos valores atípicos, y la de ciclos con crepitancias y sibilancias también (aunque en menor medida), se decidió tratarlos como distribuciones **no** normales. Al igual que el promedio del espectro para cada uno, debido a que para éste, las gráficas cuartil-cuartil evidencian que los datos no siguen la distribución Gaussiana ideal.

Por este motivo las técnicas aplicadas para las pruebas de hipótesis fueron No paramétricas; en las cuales se evidencia de la correlación de Spearman, que el **promedio móvil fino** de los ciclos sanos se relaciona en un 5.18% con el de los ciclos enfermos, y aunque es un relación bastante baja, existe un 28.40% de probabilidad de obtener el peor de los casos, y que los ciclos sean iguales, lo cual reafirma la prueba de Mannwhitneyu con su probabilidad del 33.85%, por lo que para el caso 1, no es posible aceptar la hipótesis alternativa y debemos tomar la hipótesis nula en la cual el promedio móvil fino de los ciclos sanos es **similar** al de los enfermos.

Por otro lado, para el caso 2, la prueba de Spearman arroja una correlación del 9.94% con una probabilidad del 3.9% de obtener el peor de los casos, lo cual es aceptable bajo un umbral del **5%** de error, y de forma análoga al caso anterior, la prueba de Mannwhitneyu refuerza este hecho con su probabilidad aproximadamente nula de obtener datos localizados en el peor de los casos donde son iguales; por lo que para el **promedio del espectro de potencia** se rechaza la hipótesis nula y se acepta la hipótesis alternativa en la cual está característica es **diferente** para los ciclos sanos, comparados con los que tienen crepitaciones y sibilancias.

V. Conclusión

En conclusión, la relación existente entre los ciclos respiratorios sanos, es similar cerca de un 10% con los ciclos que tienen crepitancias y sibilancias, sin embargo en un 96.1% de las ocasiones, es probable distinguirlos mediante el promedio de la potencia espectral. Y aunque en este estudio sólo se evaluaron estos 2 índices, es posible encontrar otros patrones en los demás, que permitan diferenciar y categorizar los ciclos respiratorios sanos de los enfermos.

Además, se tiene como ventaja el hecho de que ya se cuenta con un archivo de Excel que recopila toda la información tanto de los pacientes como de sus audios de auscultación, el cual es de fácil manipulación por lo menos en el software de Python que permite importarlo como un DataFrame y aplicar todas las herramientas necesarias que dispone el paquete de análisis de datos de la librería Pandas. Por lo que este estudio puede tomarse como punto de partida para encontrar aspectos característicos con miras a la automatización del proceso de clasificación e incluso de diagnóstico de enfermedades pulmonares y se deja al pendiente la continuación del notebook "**Statistics**".

VI. Anexos (**IMPORTANTE**)

El desarrollo de este estudio se gestionó en el repositorio de **GitHub** y puede encontrarse en el siguiente enlace: https://github.com/santiagortiiz/Auscultation-Signals.git

Se anexan **7 archivos .py** en los cuales se desarrolló todo el proyecto, **se recomienda leerlos en el siguiente orden** para comprender a cabalidad la metodología empleada en este estudio:

-Filter_design.py: Diseño de filtros digitales.
-Filter_routine.py: Rutina que permite crear un filtro, aplicarlo, y visualizar su comportamiento en un diagrama de Bode.
-Wavelet.py: Contiene la clase procesador que ejecuta la rutina de descomposición de una señal, filtrado wavelet, y reconstrucción por transformada de Haar.
-Processor.py: Contiene la rutina que se encarga de limpiar una señal de auscultación aplicando secuencialmente filtros digitales y el filtro wavelet, además de permitir **graficar** el análisis frecuencial y **exportar** la señal a formatos **.mat o .wav** en cualquier etapa del proceso.
-Features.py: Contiene el método features que permite calcular los índices de interés en este estudio y los retorna en forma de diccionario.
-Auscultation_signals.py: Rutina que llama el método process del archivo Processor.py, y el método features del archivo Features para generar una señal de auscultación filtrada, y extraer los índices de interés.
-Database_Processing.py: Script que aplica la rutina descrita en Auscultation_signals.py sobre cada uno de los elementos de la base de datos, y genera una tabla de Excel que recopila toda la información disponible en la base de datos trabajada, en conjunto con la información extraída del procesamiento de las señales.

Finalmente se anexa el archivo de **Excel** que generó Database_Procesing.py, y el archivo **Statistics**, el cual es un Jupyter Notebook que contiene un análisis detallado de los datos recopilados y con base en el cual se realizó este artículo.

VII. Referencias

[1] M. Grønnesby, "Automated Lung Sound Analysis," 2016.
[2] Marsh, "Respiratory Sound Database," 2019. [Online]. Available: https://www.kaggle.com/vbookshelf/respiratory-sound-database. [Accessed: 29-May-2020].