UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR

POSGRADOS

MAESTRÍA EN

ELECTRÓNICA Y AUTOMATIZACIÓN

RPC-SO-19-No.277-2018

OPCIÓN DE TITULACIÓN:

ARTÍCULOS PROFESIONALES DE ALTO NIVEL

TEMA:

CLASIFICADOR BINARIO INTELIGENTE BASADO EN REDES NEURONALES CONVOLUCIONALES PARA EL RECONOCIMIENTO DEL SONIDO DE LA TOS

AUTOR:

PAÚL ANDRÉS ANDRADE BARRIGA

DIRECTOR:

CHRISTIAN RAÚL SALAMEA PALACIOS

CUENCA - ECUADOR
2021

COHORTE
2018 - 2020

Autor:

Paúl Andrés Andrade Barriga

Ingeniero Electrónico
Candidato a Magíster en Electrónica y Automatización, Mención en Informática Industrial por la Universidad Politécnica Salesiana - Sede Cuenca. pandradeb@est.ups.edu.ec

Dirigido por:

Christian Raúl Salamea Palacios

Ingeniero Electrónico.
Máster en Diseño, Gestión y Dirección de Proyectos.
Doctor en el Programa Oficial de Doctorado en Ingeniería de Sistemas Electrónicos por la Universidad Politécnica de Madrid.
csalamea@ups.edu.ec

CLASIFICADOR BINARIO INTELIGENTE BASADO EN REDES NEURONALES CONVOLUCIONALES PARA EL RECONOCIMIENTO DEL SONIDO DE LA TOS

Clasificador binario inteligente basado en redes neuronales convolucionales para el reconocimiento del sonido de la tos.

Paul Andrade[1], Christian Salamea[1,2]
[1]*Interaction, Robotics and Automation Research Group, Universidad Politécnica Salesiana*
Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador.
pandradeb@est.ups.edu.ec
[2]*Speech Technology Group, Information and Telecommunication Center, Universidad Politécnica de Madrid*
Ciudad Universitaria Av Complutense 30, 28040, Madrid,España.
csalamea@ups.edu.ec

Resumen— En esta investigación se propone una red neuronal convolucional (CNN) como clasificador binario para identificar el sonido de la tos y aislarlo de otros sonidos tales como el habla o ruido ambiental. Se utiliza una base de datos con grabaciones en formato "WAV"que se dividen en tramas más pequeñas para llevar a cabo el procesamiento. En cada trama de audio se realiza la extracción de características y se usan los coeficientes cepstrales de MEL (MFCC's), los cuales se han sido obtenidos dedos maneras con el fin de contrastar sus resultados, primero por medio de HTK, y segundo, utilizando un procedimiento para obtener los MFCC's sin usar HTK. Los resultados obtenidos indican que con los dos métodos se obtienen resultados similares para entrenamiento, prueba y validación con la red neuronal. Dado que los datos son limitados se aplica una técnica de Aumento de Datos (Data Augmenting) conocida como „pitch shifting" para el aumento artificial de datos en los dos métodos y se evalúa en qué medida esta técnica contribuye a mejorar los porcentajes de entrenamiento, prueba y validación de la red.Los resultados obtenidos aplicando esta técnica en el primer método muestran una mejora relativa del 4 %, 7 % y 2.7 % respectivamente, y al usarlo en el segundo método muestran una mejora del 23 %, 7 % y 14 % respectivamente. Al comparar los dos métodos propuestos, antes y después de haber aplicado el aumento de datos se obtiene hasta un 4 % de mejora en la validación.

Abstract— In this research, a convolutional neural network (CNN) is proposed as a binary classifier to identify the sound of coughing and isolate it from other sounds such as speech or environmental noise. A database with recordings in "WAV"format that are divided into smaller frames is used for processing. Features extraction is carried out in each audio frame and the MEL cepstral coefficients (MFCC's) are used, which have been obtained in two ways, in order to contrast their results, first throught HTK software, and second , using a procedure to obtain the MFCC's without using HTK. The results obtained indicate that the two methods obtain similar results for training, testing and validation with the neural network. Due to the data are limited, a Data Augmenting technique known as „pitch shifting" is applied for the artificial increase of data in the two methods and it is evaluated how this technique contributes to improving the percentages of training, testing and validation of the neural network. The results obtained by applying this technique in the first method show a relative improvement of 4 %, 7 % and 2.7 % respectively, and when using it in the second method they show an improvement of 23 %, 7 % and 14 % respectively. When comparing the two proposed methods, before and after applying Data Augmenting, up to 4 % improvement in validation is obtained.

***Index Terms*—Red Neuronal Convolucional (CNN), Análisis de la tos, Aprendizaje profundo, inteligencia artificial, Espectrogramas, Coeficientes de MEL, Data Augmenting.**

I. INTRODUCCIÓN

Debido al aumento del uso de sistemas inteligentes en aplicaciones de reconocimiento de patrones, el aprendizaje automático de las máquinas viene evolucionando continuamente y cada vez con mejores resultados. Uno de los mejores diseños es la red neuronal convolucional (CNN) [1].

Esta red neuronal presenta una tecnología que combina redes neuronales artificiales y estrategias de aprendizaje profundo [2]. Originalmente fue diseñada para el análisis de imágenes, tales como espectrogramas, que son las imágenes que utilizaremos en este trabajo, pero también han demostrado tener un excelente rendimiento al procesar lenguaje natural [3], o en el reconocimiento y clasificación de sonidos ambientales, sistemas de alarmas de seguridad, etc [4].

Cuando se analizan sonidos tales como: el estornudo, la voz y la tos, que es precisamente el sonido en el cual se centra esta investigación, un sistema de red neuronal artificial (ANN) junto a un sistema de clasificación pueden ayudar de varias maneras, por ejemplo, a mejorar la calidad de vida de un paciente ayudando al médico en su diagnóstico o también a las autoridades de salud pública a tomar decisiones para disminuir costos en los servicios de salud con el uso de estos sistemas inteligentes etc [5].

La mayoría de los sistemas que se utilizan para clasificar sonidos se dividen en dos fases principales: extracciónde características y clasificación. En la fase de extracciónde características, la información contenida en la señal se transforma en características numéricas que se usan como entradas para el clasificador, en donde son reconocidas como un tipo de sonido en particular [4].

Por otra parte, la actual pandemia de COVID-19 no solo amenaza la vida, la salud y la productividad de las personas [6], sino que también ha afectado la economía de todos los países. El 30 de enero de 2020 la OMS declaró la emergencia sanitaria en todo el mundo, en tanto que el 28 de abril de ese mismo año habían más de 3 millones de infectados y no había una vacuna para prevenirlo [7].

El uso de la tecnología de la información con un enfoque en campos como la ciencia de datos y el aprendizaje automático puede ayudar en la lucha contra esta pandemia. Es importante contar con métodos de alerta temprana a través de los cuales se pueda pronosticar cuanto afectará la enfermedad a la sociedad y así el gobierno pueda tomar acciones necesarias sin afectar su economía [7].

El aprendizaje automático (Machine Learning) y el aprendizaje profundo (Deep Learning) se han utilizado entemas relacionados con el (COVID-19), principalmente ensistemas de predicción de aumento de casos o, en general, de evolución de la pandemia [8]. Una de las herramientas más utilizadas han sido las imágenes de rayos X torácicas con las que se pueden encontrar características visuales relacionadas con la enfermedad.

Estudios anteriores han analizado exámenes médicos basados en los sonidos de la tos. Abaza en [9] analizó las características del flujo de aire y el sonido de una tos saludable para entrenar a un clasificador que distinga entre sujetos sanos y aquellos con algún tipo de enfermedad pulmonar. Su modelo incorpora un algoritmo de reconstrucción que utiliza análisis de componentes principales. Obtuvo una precisión del 94 % y97 % para identificar la fisiología pulmonar anormal en sujetos femeninos y masculinos respectivamente [9].

Murata en [10] utilizó formas de onda expandidas en el tiempo combinadas con espectrogramas para diferenciarentre tos con flema y tos seca. El análisis del sonido de latos también se ha utilizado para diagnosticar la neumoníay Swarnkar en [11] lo utilizó para evaluar la gravedad del asma aguda. Este último informó que su modelo puede predecir entre niños que padecen dificultades respiratorias que involucran asma aguda y puede caracterizar la gravedad de la constricción de las vías respiratorias.

Botha en [12], investigó la detección de tuberculosis (TB) utilizando información espectral a corto plazo extraída de los sonidos de la tos. Informaron una precisión del 78 %al distinguir entre la tos de los pacientes con tuberculosispositiva y el grupo de control sano. Además, se observó que la precisión de la detección de TB aumentó al 82 % cuando se incluyeron las mediciones clínicas junto con las características extraídas del audio de la tos. Los sonidos de la tos utilizados en algunas de las investigaciones antes mencionadas se registraron cuidadosamente en entornos de estudio, mientras que la base de datos utilizada en esta investigación se recopiló utilizando un teléfono inteligente en un entorno hospitalario real.

Esta estrategia de recolección de datos tiene implicaciones profundas cuando se examinan los sonidos de tos sintomáticos asociados con COVID-19, debido a que la tos es un síntoma principal, junto con la fiebre y la fatiga. Se entrenaron sistemas basados en redes neuronales convolucionales (CNN) para detectar la tos y reportar COVID-19, se informó una precisión superior al 90 % en [13]- [14] mientras que otro estudio había informado una precisión del 75 % [15].

En otro estudio se extrajeron características de una base de datos de múltiples fuentes que contenía sonidos respiratorios y de tos [16] que se utilizaron para entrenar una máquina de vectores de apoyo (SVM) y ensamblar un clasificador para detectar individuos con COVID-19, informaron una precisión de alrededor del 80 %.

Balamurali en [17] propone un modelo de clasificación de los sonidos de la tos basado aprendizaje profundo que puede distinguir entre niños con tos sana y patológica, como asma, infección del tracto respiratorio superior e inferior, recopila un nuevo conjunto de datos de sonidos de tos, etiquetados con el diagnóstico de un médico. El modelo elegido es una red bidireccional de memoria a largo y corto plazo (BiLSTM) basada en características de coeficientes cepstrales de frecuencia Mel (MFCC). El modelo entrenado resultante cuando se entrena para clasificar dos clases de toses sanas o patológicas (en general o pertenecientes a una patología respiratoria específica) - alcanza una precisión superior al 84 % al clasificar la tos según la etiqueta proporcionada por el diagnóstico del médico.

Bansal en [18] utiliza un dataset público ESC-50 con sonidos de tos y los etiqueta manualmente en dos clases, COVID y NO COVID, con este dataset propone un clasificador basado en redes neuronales convolucionales con dos enfoques diferentes, uno basado en la extracción de características MFCC y el otro basado en imágenes de espectrogramas como datos de entrada a la red neuronal, obtiene un 78 %de precision con un 81 % de sensibilidad en el enfoque con MFCC y es mejor que el enfoque con espectrogramas.

En la siguiente investigación se propone un modelo de red neuronal convolucional (CNN) basado en la extracción de características mediante los coeficientes cepstrales de MEL (MFCC' s) a una base de datos inicial con audios en formato ".wav" que a diferencia de los estudios antes mencionados serán divididos en pequeñas tramas separando así el sonido de la tos de otros sonidos, lo que ayuda a mejorar el análisis fonotáctico de la señal. La base de datos tiene sonidos de tos de diferentes patologías como asma, infecciones en eltracto superior e inferior, tos seca, tos con flema, tos sanay también tos de pacientes con COVID-19 lo que aporta una gran variabilidad para la clasificación del sonido de la tos.

Con los coeficientes de Mel se obtienen los espectrogramas que servirán para el entrenamiento de la red neuronal convolucional (CNN), estos espectrogramas son obtenidos con dos diferentes métodos, en el primero se procesan las tramas de audios con las librerías de HTK y en el segundo se procesan las tramas de audios mediante un proceso matemáticogeneral para la obtención de MFCC's [19] - [20], luego se comparan los resultados obtenidos en los dos procedimientos y finalmente se analiza y se aplica la técnica de Aumentode Datos (Pitch Shifting) para evaluar la contribución del incremento artificial de datos en el entrenamiento de la CNN.

II. Preliminares

II-A. Análisis de la Tos

La tos se presenta por una expulsión repentina de aire de las vías respiratorias que se caracteriza por un sonido típico. Este sonido presenta características que permiten identificar a la tos y diferenciarla de otras manifestaciones vocales. Es uno de los síntomas más comunes dentro de un amplio rango de enfermedades respiratorias como la bronquitis, neumonía

y asma. El análisis del registro sonoro de la tos tiene unvalor significativo en el pronóstico de una enfermedad porque sus cambios pueden indicar la efectividad de la terapia o el progreso de la enfermedad [21].

Estudios como el realizado por Keleman [22] han demostrado que existen al menos tres fases para cualquier tos en particular, es decir, un estallido inicial debido a la turbulencia del aire y vibración del tejido, seguida de una fase ruidosa, y la última un estallido vocálico cuando la glotis corta con fuerza el flujo de aire[23].Estos estudios han llevado al desarrollo de métodos de evaluación de la tos asistidos por computadora [24], a su vez, las técnicas de análisis espectral de sonido que se han utilizado ampliamente para estudiar los sonidos pulmonares en el asma, también se han aplicado al sonido de la tos [25].

En relación a la localización de procesos patológicos en las vías aéreas, el primer estallido de tos refleja algunos procesos periféricos en la bifurcación traqueal. El segundo sonido contiene información sobre la situación en el área de la laringe. La parte intermedia entre dos sonidos da información sobre la situación en la tráquea [21].

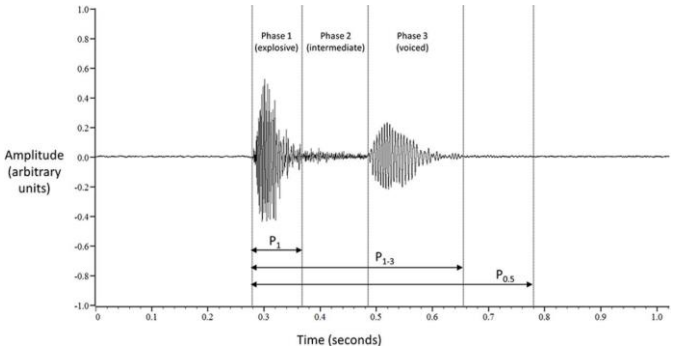

Figura 1. Fases de la Tos [26]

El análisis espectral del sonido y sus gráficos relacionados se han utilizado para analizar varios tipos de timbre del sonido de la tos. El timbre puede ser muy típico, está determinado por la frecuencia fundamental (rango entre 300 Hz a 700 Hz en condiciones normales) al igual que para otros sonidos que pueden considerarse perturbaciones de la función respiratoria generadas por reflejo tales como estornudos, hipo, suspiros, etc. [27].

Otro método de evaluación del sonido de la tos, en frecuencia, es la Transformada de Fourier (FFT), ha sido utilizada por muchos años para analizar varios tipos de sonidos como el ritmo cardíaco, la voz, la respiración y sonidos pulmonares [28].

En los casos de pacientes con enfermedades pulmonares crónicas se observa en la mayoría de los casos que en el espectrograma hay una primera fase explosiva seguida de unos rápidos cambios en la energía en distintas frecuencias, después de esta primera fase le sigue una segunda fase en donde los cambios son más suaves y se mantienen estables [28].

En el caso de pacientes con asma, la tos se caracteriza por su larga duración seguido de un silbido en una frecuencia menor [28], se compara mediante espectrogramas el sonido de la tos de niños asmáticos y saludables, se observa que al realizar ejercicio como correr al aire libre puede alterar el sonido de la tos [24] en los niños asmáticos ya que se han encontrado frecuencias más altas en los espectrogramas de este tipo de pacientes [25] pero este ejercicio no tiene relevancia en los niños saludables [29].

II-B. Análisis en HTK

HTK se basa en un conjunto de librerías y herramientas con formato en código C, estas herramientas proporcionan diversas y sofisticadas facilidades para el análisis del habla, la formación de HMM's (Modelos Ocultos de Markov), las pruebas y el análisis de resultados. Este software trabaja con funciones de distribución continuas y discretas para construir modelos complejos [30]. En este software se puede realizar la extracción de los coeficientes cepstrales de MEL (MFCC), mediante la creación de un archivo de configuración "config" (Fig. 2) en donde se deben indicar todos los parámetros a procesar.

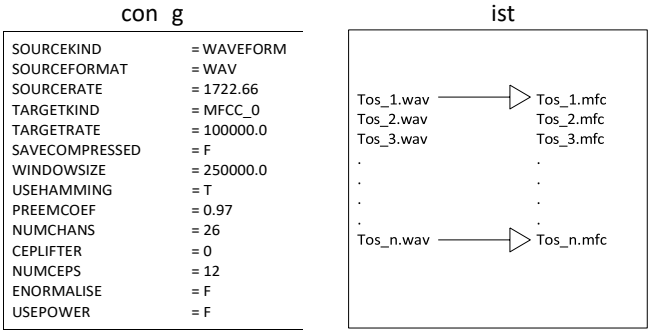

Figura 2. Archivo de Configuración y script HTK

La configuración de los parámetros son los siguientes: SOURCEKIND y TARGETKIND hacen referencia a la forma natural de los datos de entrada y en que se quieren convertir estos datos respectivamente, en este caso se requiere pasar desde una señal tipo onda a un archivo con los coeficientes MFCC's. SOURCEFORMAT sirve para especificar el formato del archivo de origen, en este caso se procesan archivos ".wav". Los parámetros SOURCERATE, TARGETRATE, WINDOWSIZE determinan la tasa de muestreo de la señal, el solapamiento (nanosegundos) y el tamaño de la ventana (nanosegundos), para audios con frecuencia de muestreo con 16Khz el parámetro SOURCERATE es de 625, para una frecuencia de muestreo de 48Khz será de 1875 y para una frecuencia de muestreo de 41Khz será 1722,6. El parámetro PREEMCOEF hace referencia al pre-énfasis que se le aplica a la señal de entrada antes de ser procesada, el valor general es de 0.97 y para atenuar los extremos de la señalen cada ventaneo que se aplica se configura el parámetro booleano USEHAMMING en T (true). NUMCHANS sirve para especificar el número de canales en el banco de filtros, NUMCEPS se utiliza para determinar el número de coeficientes cepstrales que se extraerán, para este caso se utilizan 26 y 12 respectivamente.

II-C. *Espectrogramas*

La imagen del espectrograma está en formato ".png", con un tamaño de 128x128 píxeles (Fig. 3), que serán los datos de entrada para el entrenamiento de la red neuronal. Estos espectrogramas son gráficos bidimensionales pero cuentan con una tercera dimensión que se representa en colores, el tiempo va de izquierda a derecha en el eje horizontal; el eje vertical muestra las frecuencias o tonos, las más bajas en la parte inferior del gráfico y las más altas en la parte superior. La amplitud o energía de una frecuencia está representada por la tercera dimensión, un color azul corresponde a amplitudes bajas y el color rojo corresponde a amplitudes más fuertes [31].

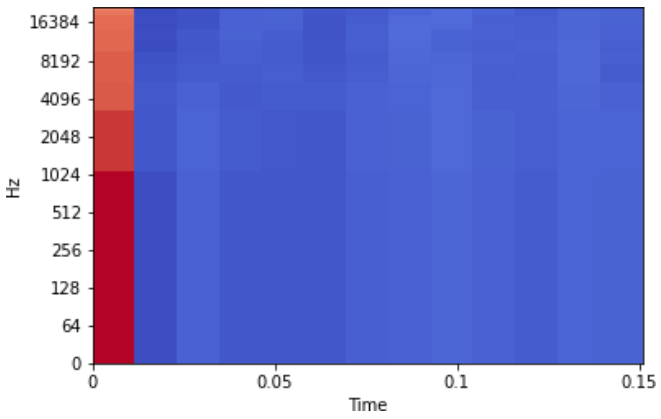

Figura 3. Espectrograma

II-D. *Red Neuronal Convolucional*

Una red neuronal convolucional esta específicamente diseñada para procesar imágenes. En su arquitectura usa una capa convolucional, una capa de agrupación (pooling) y una capa completamente conectada (fully connected) para el proceso de aprendizaje [32].

El término "convolución" hace referencia a la función matemática de convolución, en términos simples, dos imágenes que pueden ser representadas como matrices son multiplicadas para dar una salida que es usada para extraer las características de la imagen [33]. Se compone por dos grandesbloques :

II-D1. *Primer bloque*

Funciona como un extractor de características, es decir, se filtran los mapas de características obtenidos mediante los filtros de convolución, con nuevos kernel y los nuevos mapas obtenidos se normalizan o se redimensionan mediante funciones de activación y se repite el proceso hasta que los últimos valores de los mapas se concatenan en un vector, que resultara ser la entrada para el segundo bloque [34].

II-D2. *Segundo bloque*

Las combinaciones lineales y funciones de activación transforman los valores del vector de entrada para devolver un nuevo vector a la salida que contiene elementos según las clases existentes, se utiliza generalmente una función **"softmax"** como función de activación y obtener a la salida un valor similar a una distribución de probabilidad que resulta útil para la generación del modelo.

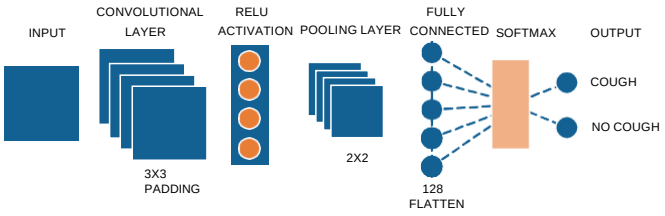

Figura 4. Arquitectura de CNN Propuesta

III. METODOLOGÍA

III-A. *Sistema General de Reconocimiento*

Para el desarrollo de esta investigación, se plantea un esquema de reconocimiento y clasificación de "Tos" y "No Tos", mediante el uso de espectrogramas los cuales se obtienen a partir del procesamiento de una base de datos de audios en formato ".wav" de los que se extraen los coeficientes cepstrales de MEL y posteriormente los espectrogramas mencionados. Los espectrogramas obtenidos se utilizan como entradas de datos para el entrenamiento de una red neuronal convolucional que funciona como clasificador y determina en su salida si existe "Tos" o "No Tos".

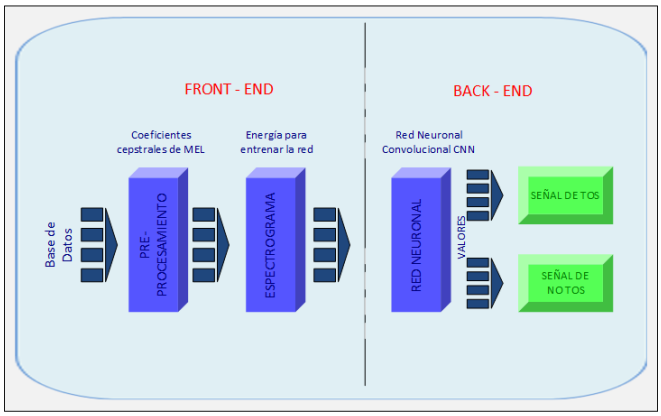

Figura 5. Esquema General de Reconocimiento

III-B. *Base de Datos*

Se tiene como punto de partida una base de datos con 266 grabaciones con sonidos de tos y 69 grabaciones con sonidos de otros eventos tales como, conversaciones, risas, o ruido ambiental. Estas grabaciones están en formato ".wav" y duran entre 1 y 112 segundos, los participantes hicieron las grabaciones de forma telemática y anónima después de aceptar las condiciones de participación expuestas en un consentimiento informado elaborado específicamente con este objetivo.

Mas adelante se hace un aumento de la base de datosinicial con bases de datos externas ya preparadas y de dominiopúblico para el entrenamiento de redes neuronales. "ESC50" [35], que contiene audios en formato ".wav" a una frecuencia

de muestreo de 44.1KHz. “COUGHVID DATASET” [36], que contiene audios en formato “.wav” a una frecuencia de muestreo de 16KHz. Posteriormente se hace un aumento de datos artificial mediante una técnica de aumento de datos denominada “Pitch Shifting”.

Cuadro I
DATASET ORIGINAL

Distribución de Audios				
Audios	Formato	Canal	Frecuencia de Muestreo (KHz)	Duración Prom.(seg)
266	.wav	Mono	48	1-112
69	.wav	Estereo	44.1	1-112

Los audios originales pueden ser separados en tramas más pequeñas aprovechando la característica de cuasi estacionalidad de señales de corta duración. Para realizar la separación y etiquetado de las tramas de audio se necesita usar un entorno supervisado, en este caso se usa el software “Audacity” el cual permite escuchar e identificar visualmente en el dominio temporal las fases de la tos expuestas en la Fig. 1.

Las tramas resultantes se etiquetan en dos clases, “Tos” y “No Tos” en el software equivalen a “1” y “0” respectivamente. En relación con la clase “Tos” tenemos únicamente el sonido de la primera fase de la tos (Fig. 1) y dentro de la clase” No Tos” tenemos la segunda fase, tercera fase y ruidos ambientales del entorno de grabación.

Las tramas obtenidas mantienen las propiedades del audio original excepto la duración ya que deben ser separadas y almacenadas como nuevos archivos que ahora se convierten en la base de datos inicial con la cual se procede a la extracción de características (Coeficientes cepstrales de MEL)mediante el software de HTK y también mediante un algoritmomatemático con la finalidad de contrastar los resultados obtenidos en estos dos casos [30][37].

La extracción de características relevantes en una señalde audio que sean robustas a variaciones es de fundamental importancia en la recuperación de información de dicha señal. En esta investigación se usan los coeficientes de Mel (MFCC’ s) debido a que son características espectrales diseñadas para el procesamiento de voz en donde las bandas de frecuencia se colocan logarítmicamente en una llamada “escala de Mel” la cual se aproxima a la respuesta del sistema auditivo humano.

III-C. Obtención de MFCC's con HTK

Después de configurar todos los parámetros para la extracción de características se coloca el archivo de configuración “config” y un archivo script “flist” (ver Fig. 2) dentro del mismo directorio de la base de datos de los audios separados por tramas. En este punto la base de datos tiene un tamaño de 3051 tramas de audios de los cuales 1787 corresponden a la clase “Tos” y 1264 corresponden a la clase “No Tos”.

Cuadro II
PRIMER DATASET

Distribución de Audios			
Audios	Formato	Canal	Frecuencia de Muestreo (KHz)
2140	.wav	Mono	48
62	.wav	Mono y Estéreo	44.1

Las facilidades para la entrada y salida de voz en HTK son proporcionadas por cinco módulos distintos ver Fig. 6. Toda entrada de voz en HTK se controla mediante parámetros de configuración que dan detalles de las operaciones de procesamiento que se deben aplicar a cada archivo de voz de entrada o fuente de audio.

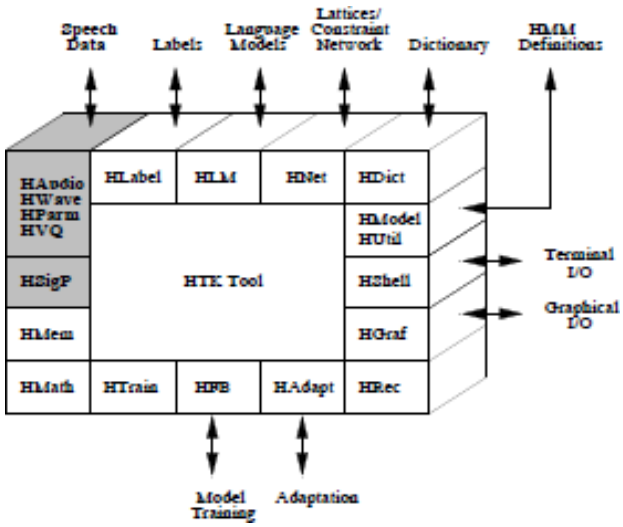

Figura 6. Módulos para Procesamiento de Voz en HTK

HTK carga los archivos “.wav” y devuelve un archivo “.mfc”. Para la obtención de los coeficientes de MEL en el script se detallan los archivos de origen “.wav” y los archivos de destino “.mfc” y se ocupa la herramienta de HTK, **HCopy** [30].(Ver Fig. 7)

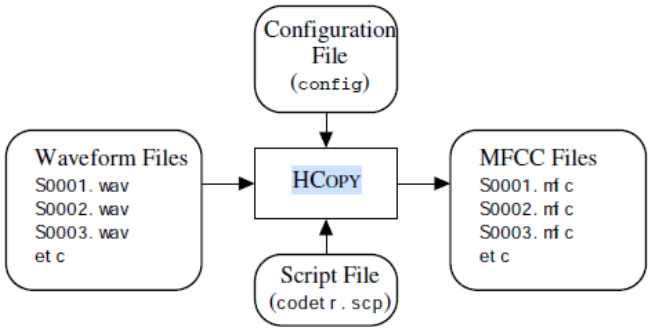

Figura 7. Herramienta de Codificación de datos

Después de configurar todos los parámetros para procesar la señal de audio la herramienta HCopy internamente realiza

primero un pre-énfasis a la señal de audio aplicando la siguiente ecuación en diferencias:

$$S'_n = s_n - k s_{n-1} \quad (1)$$

A las muestras $s_n, n = 1, N$ en cada ventana, k es el coeficiente de pre-énfasis el cual debe estar en un rango de $0 \quad k \quad 1$. En este caso se escogió un valor de 0,97. Es necesario también reducir las discontinuidades en el borde de cada ventana, esto se logra estableciendo el parámetro booleano USEHAMMING en TRUE, el cual aplica la siguiente transformación a las muestras $s_n, n = 1, N$ en la ventana:

$$S'_n = [0{,}54 \quad 0{,}46 \cos \frac{2\pi(n-1)}{N-1}] s_n \quad (2)$$

HTK proporciona un banco de filtros simple basado en la transformada de Fourier diseñado para dar aproximadamente la misma resolución en una escala de mel.Los filtros utilizados son triangulares y están igualmente espaciados a lo largo de la escala mel que se define por:

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (3)$$

Los filtros triangulares se distribuyen en todo el rango de frecuencias desde cero hasta la Frecuencia de Nyquist. Ahora los coeficientes cepstrales de mel se calculan a partir de el logaritmo de las amplitudes del banco de filtros $[m_j]$ usando la Transformada de coseno discreta.

$$C_i = \sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j - 0{,}5)) \quad (4)$$

El numero de coeficientes cepstrales y el numero de bancos de filtros N se define en el archivo de configuración.

III-D. Obtención de MFCC's sin HTK

El software de HTK internamente realiza todo el proceso matemático que generalmente se suele utilizar en caso dequerer obtener los coeficientes de MEL paso a paso. El procedimiento de manera general se describe así:[19]-[20]

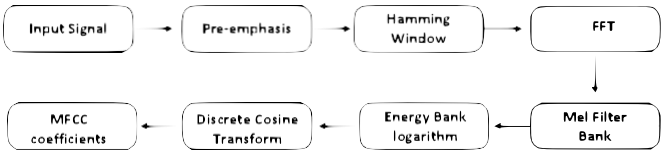

Figura 8. Extracción de MFCC's sin HTK [19]

Primero se realiza la división de la señal de audio en tramas de Tos y No Tos con el software Audacity, se carga el archivo .wav en un array de (nxm), donde "n.es el numero de tramas y "m.es el numero de muestras

tiempo se considera casi estacionaria y la información más relevante está en el centro ya que los extremos son información superpuesta de las demás tramas.

Ahora se puede realizar el cálculo de la potencia de trama, aquí se realiza un "padding" a la matriz para poder aplicar la transformada de Fourier. Su tamaño es una potencia de dos, y debe ser mayor o igual al número de muestras de la trama, luego se toma el valor absoluto de la mitad de todo el espectro ya que la otra mitad es igual a la primera y se calcula la potencia con la siguiente ecuación:

$$P_i(k) = \frac{1}{N} |S_i(k)|^2 \quad (5)$$

En donde:

P → Es la potencia de cada trama
Si(k) → espectro frecuencial de cada trama
N → Numero de muestras de cada trama

Posteriormente se crean los Bancos de Filtros de MEL, estos bancos son triangulares espaciados de manera equidistante, con amplitud máxima de 1 y para el reconocimiento de voz se consideran 26 filtros [20]. El umbral de bajas frecuencias se fija en cero y el de altas frecuencias se considera con la división de la frecuencia de muestreo entre dos, se transforma de Hz a Mel utilizando la ecuación:

$$mel(f) = 1127 \ln(1 + \frac{f}{700}) \quad (6)$$

En donde:
f→ Representa la frecuencia física para el oído humano en Hz.
mel(f) Representa la frecuencia percibida en escala de mel.

Debido a que se va a trabajar con la energía de la señal es necesario que los 26 puntos de frecuencia se aproximen a los puntos FFT que fueron calculados anteriormente, se deben escalar los triángulos en el dominio de la frecuencia utilizando una función inversa que convierte de Mel a Hz con la siguiente ecuación:

$$Hz(m) = 700(e^{\frac{m}{1127}} - 1) \quad (7)$$

En donde:

Luego se aplica un Pre-énfasis al archivo de audio medianteun filtro pasa alto, con esto se mejoran las componentes espectrales y se distribuyen de manera uniforme, se usa un valor de 0,97 [20].

Ahora se realiza el ventaneo de Hamming a cada trama,la longitud de cada ventana es de 25ms con un solapamientode 10ms [20], entendiendo que la señal en este periodo de

Hz(m) Representa la conversión de la escala de mel a una escala en Hz. →

m Representa el valor de la banda de frecuencias en escala de mel →

Después se utilizan los filtros de MEL para realizar el cálculo de energía, en una matriz de [nx26] se almacenan los coeficientes resultantes del producto punto entre el banco de filtros de Mel y la potencia de la trama que corresponde acada filtro.

Se procede luego con el cálculo del logaritmo de las tramasde energía en donde se le extrae el logaritmo natural a cada elemento de la matriz ya que el oído humano no percibe el sonido en una escala lineal, mediante el logaritmo se puede

usar una técnica de normalización de canales conocida como resta media cepstral.

Por último, se realiza el cálculo de la transformada coseno discreta, aquí para el procesamiento de la voz se utilizan los primeros 13 coeficientes cepstrales de esta transformada [20], la cual contiene 26 puntos de frecuencia que fueron logaritmizados previamente.

Al aplicar esta transformada a cada fila de la matriz, se obtiene una nueva matriz de MFCC's de [nx13], donde n representa cada una de las tramas de la señal de audio, con la siguiente ecuación:

$$C_i = \sqrt{\frac{2}{26}} \sum_{k=0}^{25} l_k \cos\left(\frac{i\pi(k + 0,5)}{26}\right) \quad (8)$$

En donde:

l → Representan las filas de la matriz de logaritmos.

C(i) Representa el índice del coeficiente de la transformada coseno que se quiere obtener.

k Representa el índice del coeficiente de la fila l en la matriz de logaritmos

IV. Resultados

Con la finalidad de obtener una métrica más estable y objetiva después de haber realizado los experimentos, se ha utilizado la técnica de validación cruzada, la cual es un método estadístico para evaluar y comparar algoritmos de aprendizaje dividiendo los datos en dos segmentos: uno se usa para entrenar y el otro se usa para validar [38].

Para la ejecución de los experimentos se utilizó la arquitectura de red planteada en la Fig. 4, en la primera fase de experimentación se usan los espectrogramas obtenidosdesde los archivos ".mfc" en HTK y en la segunda fase de experimentación se usan los espectrogramas generados sin HTK mediante el proceso matemático descrito en la sección III-D, con el fin de poder contrastar los resultados obtenidos.

IV-A. Aumento de Datos

Debido a que uno de los aspectos clave en el buen funcionamiento de las CNN's es una base de datos de grandes dimensiones, es necesario hacer un aumento artificial de nuestra base de datos original para poder lograr un mejor rendimiento de la red.

Con la adición de las bases de datos mencionadas en la sección III-B , se han obtenido 8808 tramas de audio para procesar pero debido a la poca diversidad de los datos ha sido necesario aplicar una de las técnicas de Aumento de Datos conocida como "shift pitching". La idea detrás de esta técnica es aumentar el número de muestras mediante la transposición del tono del sonido de tos (Fig. 9), en esta investigación fue necesario usar un semitono más alto y un semitono más bajo, hacerlo más alto o más bajo provocaría que se alteren las características propias del sonido de la tos [39].

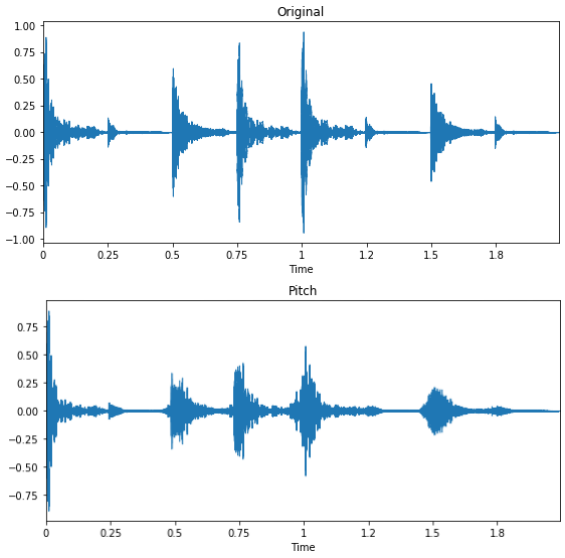

Figura 9. Pitch Shifting

Con esta técnica se ha logrado el incremento artificial de la base de datos hasta 21276 tramas de audio.

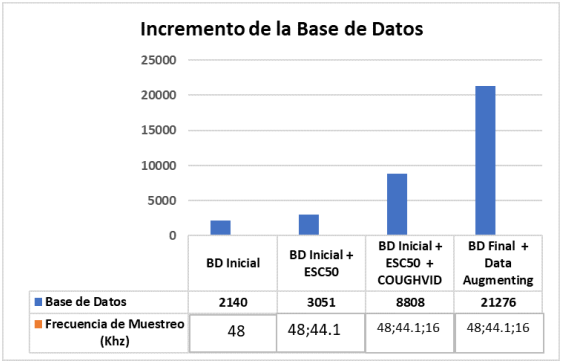

Figura 10. Incremento de la Base de Datos

IV-B. Experimentos con HTK

Con las bases de datos que han sido procesadas en HTK se han obtenido los resultados que se muestran en la Fig. 11 y al aplicar la técnica de Aumento de Datos se han obtenido los resultados expuestos en la Fig.12.

En la Fig. 11 se observan los promedios de los resultados obtenidos en entrenamiento, prueba y validación usando la base de datos inicial y la base de datos aumentada, se observa una mejora del 15 %, 13 % y 40 % respectivamente.

En la Fig. 12 se observan los promedios de los resultados obtenidos en entrenamiento, prueba y validación usando técnica de Aumento de Datos "pitch shifting" y se observa una mejora del 4 %, 7 % y 2.7 % respectivamente.

	Promedio de K-Folds	
	Database = 2140 ; k = 4 ; Epochs = 5; Batch Size = 50	
	Accuracy	Loss
Train	0.7425	0.53
Test	0.66	0.54
Validation	0.4225	0.6725
	Promedio de K-Folds	
	Database = 3051 ; k = 4 ; Epochs = 5; Batch Size = 50	
	Accuracy	Loss
Train	0.885	0.5725
Test	0.76	0.6775
Validation	0.705	0.41

Figura 11. Experimento 1

	Promedio de K-Folds con Data Augmenting	
	Database = 8808 ; k = 4 ; Epochs = 5; Batch Size = 50	
	Accuracy	Loss
Train	0.905	0.5475
Test	0.7925	0.6475
Validation	0.655	0.565
	Promedio de K-Folds con Data Augmenting	
	Database = 21276 ; k = 4 ; Epochs = 5; Batch Size = 50	
	Accuracy	Loss
Train	0.92	0.355
Test	0.82	0.4225
Validation	0.725	0.3575

Figura 12. Experimento 1 con Aumento de Datos

IV-C. Experimentos sin HTK

Con la base de datos que se ha obtenido aplicando el procedimiento matemático sin usar HTK se obtienen los resultados expuestos en la Fig. 13 y al aplicar la técnica de Aumento de Datos se han obtenido los resultados expuestos en la Fig.14

	Database = 2140 ; k = 4 ; Epochs = 5; Batch Size = 50	
	Promedio de K-Folds	
	Accuracy	Loss
Train	0.6725	0.56
Test	0.8	0.615
Validation	0.705	0.4475
	Database = 3051 ; k = 4 ; Epochs = 5; Batch Size = 50	
	Promedio de K-Folds	
	Accuracy	Loss
Train	0.7725	0.4225
Test	0.7725	0.325
Validation	0.595	0.485

Figura 13. Experimento 2

En la Fig.13 se observan los promedios de los resultados obtenidos en entrenamiento, prueba y validación usando la base de datos inicial y la base de datos aumentada se observa una mejora del 13 % únicamente en el entrenamiento.

	Database = 8808 ; k = 4 ; Epochs = 5; Batch Size = 50	
	Promedio de K-Folds con Data Augmenting	
	Accuracy	Loss
Train	0.8475	0.2575
Test	0.7725	0.35
Validation	0.7725	0.325
	Database = 21276 ; k = 4 ; Epochs = 5; Batch Size = 50	
	Promedio de K-Folds con Data Augmenting	
	Accuracy	Loss
Train	0.8875	0.2675
Test	0.835	0.345
Validation	0.6925	0.34

Figura 14. Experimento 2 con Aumento de Datos

En la Fig. 14 se observan los promedios de los resultados obtenidos en entrenamiento, prueba y validación usando técnica de Aumento de Datos “pitch shifting” y se observa una mejora del 23 %, 7 % y 14 % respectivamente.

IV-D. Resultados de HTK + Aumento de Datos

Ahora se pueden comparar los resultados obtenidos del promedio de cada K-Fold con las bases de datos iniciales procesadas desde HTK y aplicando la técnica de Aumento de Datos. En el gráfico de la Fig. 15 se observa una mejora tanto en el entrenamiento, como en el test y validación.

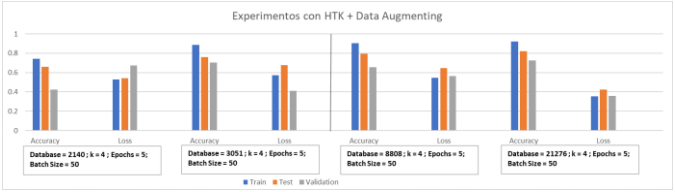

Figura 15. Experimento HTK + Aumento de datos

IV-E. Resultados Sin HTK + Aumento de Datos

Ahora se pueden comparar los resultados obtenidos en cada K-Fold, con las bases de datos iniciales obtenidas desarrollando el procedimiento matemático sin usar HTK y posteriormente aplicando la técnica de Aumento de Datos“pitch shifting”. Como se observan en las Figuras 15 y 16los resultados en el entrenamiento, test y validación mejoranen un 20 %, 20 % y 41 % respectivamente, conforme la base de datos inicial va aumentando con las bases de datos adicionalesya mencionadas y aplicando también la técnica de Aumentode Datos.

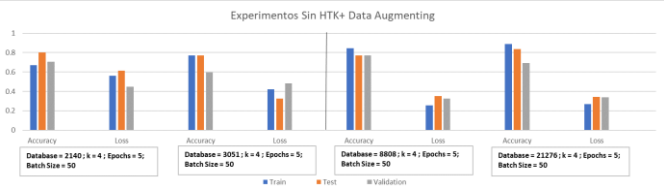

Figura 16. Experimento Sin HTK + Aumento de Datos

Al comparar los resultados obtenidos después de aplicar la técnica del Aumento de Datos tanto en la base de datos procesada en HTK como en la obtenida sin usar HTK se observan resultados similares en el entrenamiento, test y validación incluso cuando los espectrogramas en cada base de datos difieren en la colorimetría (matriz RGB).

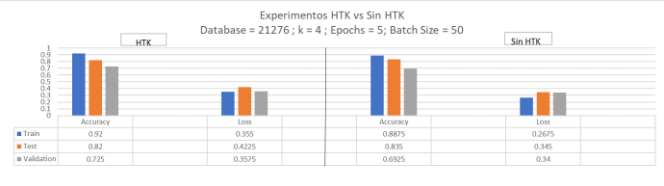

Figura 17. Experimentos con HTK y Sin HTK

V. Conclusiones

Los resultados obtenidos con todas las bases de datos Figs. 11 - 14, muestran una mejora relativa del 20 % en el entrenamiento, 19 % en el test y 41 % en la validación cuandose aplica la técnica de Aumento de Datos lo cual implica que es importante tener una base datos grande y diversa.

Fue necesario realizar un aumento de un 89 % en la base de datos inicial para obtener los resultados expuestos en la Fig. 17, de este 89 % aproximadamente un 75 % fue realizadode manera natural, es decir, añadiendo más tramas de audios desde las bases de datos ESC50 - COUGHVID y un 14 % de manera artificial, es decir, con Aumento de Datos mediante el pitch shifting con lo cual se obtuvieron las 21276 tramas de audios mencionadas en la Fig 10.

En conclusión, se obtuvieron resultados similares alprocesar la base de datos final mediante las librerías de HTK ymediante el procedimiento matemático general, a pesar de que la colorimetría de los espectrogramas en cada procedimientoes diferente, por las distintas ecuaciones matemáticas que se emplean para cada etapa, existe apenas una diferencia de 4 % a favor del uso de HTK.

La separación de los audios en tramas de Tos y No Tos brinda un mejor análisis fonotáctico de la señal de tos pero hace indispensable manejar bases de datos grandes para poder brindar información relevante suficiente a la red neuronal convolucional, por lo cual hacer uso de distintas técnicas de aumento de datos es necesario para incrementar el rendimiento de la red neuronal.

Para determinar la arquitectura óptima de la red neuronal con la que se realizaron todos los experimentos, ha sido necesario variar algunos de sus parámetros, siendo el más importante, el aumento de neuronas en la capa “fully connected”, lo cual implica un coste computacional mucho mayor pero en relación a la variación de otros parámetros de la red, con esta arquitectura se consiguen los mejores resultados.

Para una futura línea de trabajo se puede considerar al sonido de la tos como una secuencia, es decir, obtener los espectrogramas no solo de la primera fase del sonido de la tos, sino del sonido de las toses que se den entre dos inhalaciones de aire, como las que están expuestas en la Fig. 1.

En otra línea de trabajo se puede considerar la variación de algunos parámetros que son relevantes al momento de realizar la extracción de características como el tamaño de la ventana de Hamming y el tiempo de solapamiento de la señal o el número de coeficientes cepstrales que se consideran para la obtención de los espectrogramas.

VI. Agradecimientos

A la Corporación Ecuatoriana para el Desarrollo de la Investigación y Academia, CEDIA, por el financiamiento brindado a la investigación, desarrollo e innovación a través del CEPRA-XV-2021-011: Caracterización de la tos provocadapor el COVID-19 en pacientes con diagnóstico positivo. Los autores agradecen a la Escuela Politécnica Nacional, la Universidad Politécnica Salesiana y la Pontificia Universidad Católica del Ecuador.

Referencias

[1] S. Sakib, M. A. B. Siddique, and M. A. Rahman, “Performance evaluation of t-SNE and MDS dimensionality reduction techniques with KNN, ENN and SVM classifiers,” *arXiv*, 2020.

[2] S. Sakib, Ahmed, A. Jawad, J. Kabir, and H. Ahmed, “An Overview of Convolutional Neural Network: Its Architecture and Applications,” *ResearchGate*, no. November, 2018.

[3] W. Zhu, Y. Ma, Y. Zhou, M. Benton, and J. Romagnoli, *Deep Learning Based Soft Sensor and Its Application on a Pyrolysis Reactor for Compositions Predictions of Gas Phase Components*, vol. 44. Elsevier Masson SAS, 2018.

[4] P. Khunarsal, C. Lursinsap, and T. Raicharoen, “Very short time environmental sound classification based on spectrogram pattern matching,” *Information Sciences*, vol. 243, pp. 57–74, 2013.

[5] A. Teyhouee and N. D. Osgood, “Cough detection using hidden markov models,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11549 LNCS, pp. 266–276, 2019.

[6] M. A. Salam, S. Taha, and M. Ramadan, “COVID-19 detection using federated machine learning,” *PLoS ONE*, vol. 16, no. 6 June, pp. 1–25, 2021.

[7] D. Painuli, D. Mishra, S. Bhardwaj, and M. Aggarwal, “Forecast and prediction of COVID-19 using machine learning,” *Data Science for COVID-19*, pp. 381–397, 1 2021.

[8] T. Alafif, A. M. Tehame, S. Bajaba, A. Barnawi, and S. Zia, “Machine and deep learning towards covid-19 diagnosis and treatment: Survey, challenges, and future directions,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, pp. 1–24, 2021.

[9] A. A. Abaza, J. B. Day, J. S. Reynolds, A. M. Mahmoud, W. T. Goldsmith, W. G. McKinney, E. L. Petsonk, and D. G. Frazer, “Classification of voluntary cough sound and airflow patterns for detecting abnormal pulmonary function,” *Cough*, vol. 5, no. 1, pp. 1–12, 2009.

[10] A. Murata, Y. Taniguchi, Y. Hashimoto, Y. Kaneko, Y. Takasaki, and S. Kudoh, “Discrimination of Productive and Non-Productive Cough by Sound Analysis,” *Internal Medicine*, vol. 37, no. 9, pp. 732–735, 1998.

[11] V. Swarnkar, U. Abeyratne, J. Tan, T. W. Ng, J. M. Brisbane, J. Choveaux, and P. Porter, “Stratifying asthma severity in children using cough sound analytic technology,” *Journal of Asthma*, vol. 58, no. 2, pp. 160–169, 2021.

[12] G. H. R. Botha, G. Theron, R. M. Warren, M. Klopper, K. Dheda, P. D. van Helden, and T. R. Niesler, "Detection of tuberculosis by automatic cough sound analysis," *Physiological Measurement*, vol. 39, no. 4, p. 45005, 2018.

[13] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020.

[14] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[15] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, and R. Panicker, "Cough Against COVID: Evidence of COVID-19 Signature in Cough Sounds," no. Ml, 2020.

[16] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3474–3484, 2020.

[17] B. T. Balamurali, H. I. Hee, S. Kapoor, O. H. Teoh, S. S. Teng, K. P. Lee, D. Herremans, and J. M. Chen, "Deep neural network-based respiratory pathology classification using cough sounds," *Sensors*, vol. 21, no. 16, pp. 1–18, 2021.

[18] V. Bansal, G. Pahwa, and N. Kannan, "Cough classification for COVID-19 based on audio mfcc features using convolutional neural networks," *2020 IEEE International Conference on Computing, Power and Communication Technologies, GUCON 2020*, pp. 604–608, 2020.

[19] Christian Salamea Palacios Ph.D, "Coeficientes Cepstrales de Mel (MFCCs) ," 10 2019.

[20] K. S. Rao and M. K.E., "MFCC Features," *Speech Recognition Using Articulatory and Excitation Source Features*, pp. 85–88, 2017.

[21] J. Korpáš, J. Sadloňová, and M. Vrabec, "Analysis of the cough sound: An overview," *Pulmonary Pharmacology*, vol. 9, no. 5-6, pp. 261–268, 1996.

[22] S. A. Kelemen, T. Cseri, and I. Marozsan, "Information obtained from tussigrams and the possibilities of their application in medical practice," *Bulletin europeen de physiopathologie respiratoire*, vol. 23 Suppl 1, p. 51s—56s, 1987.

[23] K. P. DAWSON, C. W. THORPE, and L. J. TOOP, "The spectral analysis of cough sounds in childhood respiratory illness," *Journal of Paediatricsand Child Health*, vol. 27, no. 1, pp. 4–6, 1991.

[24] P. Piirila and A. R. Sovijarvi, "Differences in acoustic and dynamic characteristics of spontaneous cough in pulmonary diseases," *Chest*,vol. 96, no. 1, pp. 46–53, 1989.

[25] L. J. Toop, C. W. Thorpe, and R. Frightt, "Cough sound analysis: A new tool for the diagnosis of asthma?," *Family Practice*, vol. 6, no. 2, pp. 83–85, 1989.

[26] K. K. Lee, S. Matos, K. Ward, G. F. Rafferty, J. Moxham, D. H. Evans, and S. S. Birring, "Sound: A non-invasive measure of cough intensity," *BMJ Open Respiratory Research*, vol. 4, no. 1, pp. 1–9, 2017.

[27] M. Scoble, "Book Review: Book Review," *Systematic Entomology*, vol. 30, no. 3, pp. 497–498, 2005.

[28] L. Debreczeni, J. Korpas, and D. Salat, "Spectral analysis of cough sounds recorded with and without nose clip," *Bulletin européen de physiopathologie respiratoire*, vol. 23 Suppl 1, pp. 57s–61s, 1987.

[29] P. M. Olia, P. Sestini, and M. Vagliasindi, "Acoustic parameters of voluntary cough in healthy non-smoking subjects," *Respirology*, vol. 5, no. 3, pp. 271–275, 2000.

[30] R. Crichton, "Speech Input/Output," *Electronics and Power*, vol. 32, no. 9, p. 680, 1986.

[31] Vande Kamp Jade, "Spectrogram - Signal Analysis - Vibration Research," 5 2020.

[32] M. Momeny, M. A. Sarram, A. M. Latif, R. Sheikhpour, and Y. D. Zhang, "A Noise Robust Convolutional Neural Network for Image Classification," *Results in Engineering*, vol. 10, no. April, p. 100225, 2021.

[33] MK Gurucharan, "Basic CNN Architecture: Explaining 5 Layers of Convolutional Neural Network," 12 2020.

[34] Kousai Smeda MSc, "Understand the architecture of CNN | | Towards Data Science," 10 2019.

[35] K. J. Piczak, "ESC: Dataset for environmental sound classification," *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*,pp. 1015–1018, 2015.

[36] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," pp. 1–11, 2020.

[37] J. W. Duncan, "The fundamentals of...," *Aircraft Engineering*, pp. 16–20, 1945.

[38] A. Bhattacharya, "Curse of Dimensionality," *Fundamentals of Database Indexing and Searching*, pp. 141–148, 2014.

[39] L. Rafael Aguiar, M. G. Yandre Costa, and N. Carlos Silla, "Exploring Data Augmentation to Improve Music Genre Classification with ConvNets," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2018-July, pp. 1–8, 2018.