UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE CUENCA
CARRERA DE COMPUTACIÓN

DISE. O Y DESARROLLO DE UN SISTEMA PROTOTIPO PARA RECONOCIMIENTO AUTOMÁTICO DEL HABLANTE EMPLEANDO TÉCNICAS DE APRENDIZAJE PROFUNDO

Trabajo de titulación previo a la obtención del
título de Ingeniero en Ciencias de la Computación

AUTOR: JOSÉ ESTEBAN CALLE CHUCHUCA
TUTOR: ING. VLADIMIR ESPARTACO ROBLES BYKBAEV, Ph.D.

Cuenca - Ecuador
2022

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Yo, José Esteban Calle Chuchuca con documento de identificación N° 0107166555 manifiesto que:

Soy el autor y responsable del presente trabajo; y, autorizo a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Cuenca, 07 de marzo del 2022

Atentamente,

_____

José Esteban Calle Chuchuca
0107166555

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Yo, José Esteban Calle Chuchuca con documento de identificación Nº 0107166555, expreso mi voluntad y por medio del presente documento cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autor del Artículo Académico: "Diseño y desarrollo de un sistema prototipo para reconocimiento automático del hablante empleando técnicas de aprendizaje profundo", el cual ha sido desarrollado para optar por el título de: Ingeniero en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, 07 de marzo del 2022

Atentamente,

_____

José Esteban Calle Chuchuca
0107166555

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Vladimir Espartaco Robles Bykbaev con documento de identificación N° 0300991817, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: DISEÑO Y DESARROLLO DE UN SISTEMA PROTOTIPO PARA RECONOCIMIENTO AUTOMÁTICO DEL HABLANTE EMPLEANDO TÉCNICAS DE APRENDIZAJE PROFUNDO, realizado por José Esteban Calle Chuchuca con documento de identificación N° 0107166555, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Cuenca, 07 de marzo del 2022

Atentamente,

Ing. Vladimir Espartaco Robles Bykbaev, Ph.D.
0300991817

DEDICATORIA

Dedico este trabajo a mis padres Gladys Chuchuca y Juan Calle, gracias por ser el pilar y apoyo en todos estos años de carrera.

A mis hermanos Mayra Calle y Juan Fernando Calle que siempre vieron en mi un ejemplo de dedicación y constancia.

A mis amigos mas cercanos que estuvieron siempre empujándome a lo largo de la carrera para no rendirme.

José Esteban Calle

AGRADECIMIENTOS

Agradezco a mis padres y hermanos por el cariño, apoyo y paciencia a lo largo de todos estos años.

Agradezco a mi tutor de tesis por darme esa esperanza de poder finalizar este proyecto a pesar de todas las complicaciones que se presentaron en el camino.

Agradezco a mis amigos cercanos que siempre estuvieron pendientes de que llegue a la meta y siga adelante sin rendirme, a pesar de que, por momentos, hubiese parecido que no había luz en el camino.

José Esteban Calle

RESUMEN

Los sistemas de reconocimiento de voz automático del habla se encuentran dentro de un campo con una amplia capacidad de ser explotada más allá de solo ser utilizada para ordenes y comandos para ciertas actividades simples. En el ámbito de seguridad existen varios proyectos en los cuales este sistema es usado para permitir accesos, como a una vivienda o archivos, pero en si la problemática que se presenta va un poco más allá.

Los sistemas de seguridad que presentan reconocimiento de voz están basados en reconocimiento de texto e identificación de frecuencias naturales de un ser humano, pero no desarrollados a profundidad. Los proyectos más destacables que se pueden encontrar con reconocimiento de voz van de la mano de la domótica para acceso a viviendas o módulos de comandos de voz que puedan servir a personas con discapacidad a cumplir ciertas acciones para las cuales están inhabilitadas.

En la actualidad y con una gran cantidad de tecnología que simula voces humanas, cada vez es más difícil reconocer si nos encontramos hablando con un ser humano o con una inteligencia artificial, por lo que este avance tecnológico puede ser aprovechado tanto de una manera correcta como no. Este proyecto pretende sentar las bases de un sistema con un potencial a desarrollar que pueda ser utilizado para un nivel mas alto de seguridad basándonos en características esenciales de la voz humana, como la frecuencia fundamental (F0) sus armónicos o frecuencias formantes, Coeficiente de frecuencias ceptrales de Mel (MFCC), el cual muestra características asociadas al tracto vocal, entre muchas otros parámetros que pueden ser extraído con el simple análisis de una espectro de frecuencia emitida por la voz humana. Todas estas características se proponen ser aplicadas con redes neuronales profundas que aprendan de la naturaleza de la voz humana y puedan ser capaces de identificar a individuos en especifico como un sistema biométrico de voz.

Palabras clave: Reconocimiento de voz, LIBROSA, Frecuencia fundamental (F0), Mel, Hercios (Hz), CNN, STFT, MFCCS, espectrograma.

ABSTRACT

Automatic speech recognition systems are a field with a wide capacity to be exploited beyond just being used for orders and commands for certain simple activities. In the field of security there are several projects in which this system is used to allow access to files por example, but the problem is a bit more complex.

Security systems featuring voice recognition are based on text recognition and natural frequency identification of a human being, but not in deep development. The most important projects that can be found with voice recognition are, home automation for access to homes or voice command modules that can serve people with disabilities to carry out certain actions for which they are disabled.

At present and with a large amount of technology that simulates human voices, it is increasingly difficult to recognize if we are talking to a human or to an artificial intelligence. This technological advance can be used correctly or not. This project aims to lay the foundations for a system with a potential to develop that can be used for a higher level of security based on essential characteristics of the human voice, such as the fundamental frequency (F0) its harmonics or formant frequencies, Mel Frequency Ceptral Coefficients (MFCC), which shows characteristics associated with vocal treatment, among many other parameters that can be extracted with the simple analysis of a frequency spectrum emitted by the human voice. All these features are proposed to be applied with deep neural networks that learn from the nature of the human voice and may be able to identify specific individuals as a biometric voice system.

Keywords: Voice Recognition, LIBROSA, Fundamental Frequency (F0), Mel, Hertz (Hz), CNN, STFT, MFCCS, spectrogram.

INDICE DE CONTENIDO

I. Introducción

Los sistemas de reconocimiento de voz forman parte de un gran conglomerado de sistemas biométricos que interactúan con el usuario hoy en día. Según (Barrios, 2018) un sistema de reconocimiento de voz "es la capacidad que presenta un ordenador para recibir los datos de voz de un usuario, transformar la señal en código binario, el cual es asimilado por la computadora para luego establecer la comunicación hombre-máquina".

La voz cuenta con una gran variedad de características que ayudan a identificar a cada una de las personas, siendo esta una modalidad de biometría que puede ser fuertemente desarrollada. Dentro de las características tenemos la frecuencia, amplitud, la longitud de la onda y el timbre generados por la voz.

1. Frecuencia

Las frecuencias son oscilaciones dadas en un tiempo determinado, dependiendo de la forma de esta, podemos determinar si los sonidos emitidos, en este caso, por una voz humana son graves o agudos. Estas frecuencias tienen una unidad de medida denominada hercios (Hz) y 1 Hz es equivalente a un ciclo de compresión y descompresión de onda por segundo.

a. Fundamental (Fo)

La voz es una onda que contiene varios parámetros que se puede extraer y caracterizarla. Una de estas características mas utilizada es la frecuencia fundamental (F0). Esta frecuencia esta estrechamente relacionada con el "pitch" que los seres humanos percibimos de las otras voces. También esta frecuencia es muy útil para identificar la vibración de las cuerdas vocales. Esta característica en especifico es de gran utilidad a la aplicación de reconocimiento de voz, pero lastimosamente no se han encontrado muchos estudios que estudien a profundidad o que apliquen **Fo** como un eje central para un método de reconocimiento de voz dependiendo del locutor. Las frecuencias fundamentales de un hombre adulto oscilan entre los 85 y 180 Hz, mientras que la de una mujer adulta esta entre los 165 a 255 Hz

2. Longitud de onda

La longitud de onda se refiere a la distancia que se encuentra entre dos puntos a partir de donde esta ser repite, esta se identifica con la letra griega “Lambda”.

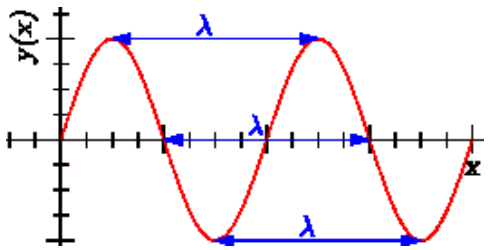

Figure 1: Longitud de onda de una señal senoidal

3. Timbre

El timbre vocal hace referencia al espectro especifico de una voz. Esta formado por formantes vocales que conjuntamente forman los armónicos que es lo que da la identidad a la voz.

a. Formantes de la voz

Los formantes vocales son los picos que se generan en el espectro del armónico del sonido emitidos por la voz, estos están asociados, pero no de manera estricta. A partir de la frecuencia fundamental nacen estos formantes siendo estos sus armónicos. Estos formantes son esenciales para el reconocimiento del habla ya que forman parte de las frecuencias predominantes en cada voz

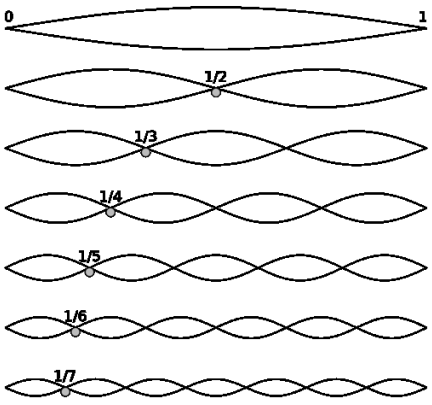

Figure 2: F0 y sus armónicos

4. Coeficiente de frecuencias cepstrales de Mel y la Escala de Mel

También conocido con las siglas MFCC son coeficientes que se basan en la percepción de la audición humana. Estas muestran características asociadas al tracto vocal.

Los coeficientes se derivan de la Transformada de Fourier o de la transformada del coseno discreta; la peculiaridad de MFCC es que lo intervalos de frecuencia que se encuentran por encima de los 500 Hz no son percibidos por los seres humanos de manera exponencial, si no de manera lineal, el odio humano esta adaptado para escuchar de mejor manera frecuencias graves que agudas, siendo así, intervalos de frecuencias superiores a los 500 Hz son reducidos dos octavas en la escala de Mel. De esta forma se adapta la información y se simula la respuesta auditiva de los seres humanos. (Rincón, C.,2007). En base a esto se creo una unidad de medida de tono llamada mel que transforma, mediante operaciones matemáticas, las señales inaudibles en Hz a señales igualmente distantes en mels, creándose así la escala de Mel.

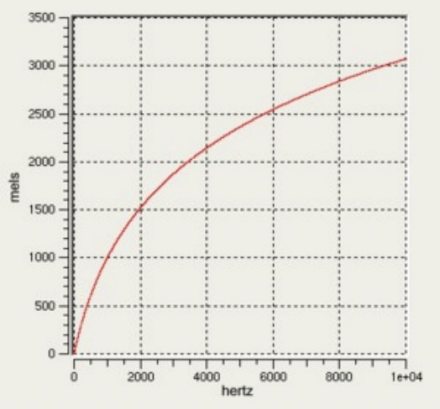

Figure 3: Escala de Mel

5. Transformada de Fourier

Mediante este calculo matemático, nos permite analizar el espectro de una señal, pasando de un espacio de tiempo discreto a un espacio de frecuencias donde podemos extraer las frecuencias relevantes de la señal. Al aplicar la transformada de Fourier a una señal, en este caso un audio, obtendremos los coeficientes de las funciones senoidales que componen a la señal original.

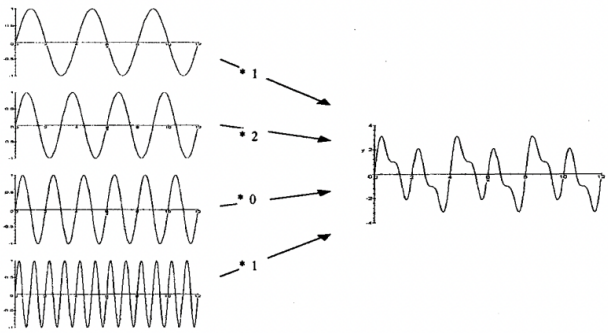

Figure 4: Señal principal descompuesta en sus coeficientes senoidales

6. CNN (Red Neuronal Convolucional)

Red neuronal artificial donde las neuronas son campos que receptan la información de una manera parecida a como un cerebro lo haría en la corteza visual. En el campo auditivo nos es de utilidad al utilizar MFCC y generar información aplicable a un espectrograma, de esta manera clasificaría la información al igual que los pixeles de una imagen. En este tipo de redes neuronales contamos la capa de entrada; capas de convolución en la cual, mediante operaciones matemáticas de sumas y multiplicaciones, en base a una cierta cantidad de filtro que genera un mapa de características, dichas características correspondientes a una posible ubicación del filtrado de la imagen original, en nuestro caso, las características comunes de un audio ingresado en la capa de partida. Mediante las capas de reducción se disminuye los parámetros para quedarnos con las características mas comunes. Todas estas características son extraídas mediante operaciones de promedios o máximos. Y Finalmente tenemos las capas de clasificación en la cual tenemos tanta cantidad de neuronas como pixeles, o en nuestro caso señales de frecuencia, que se conectan entre si mediante diferentes capas, capas ocultas.

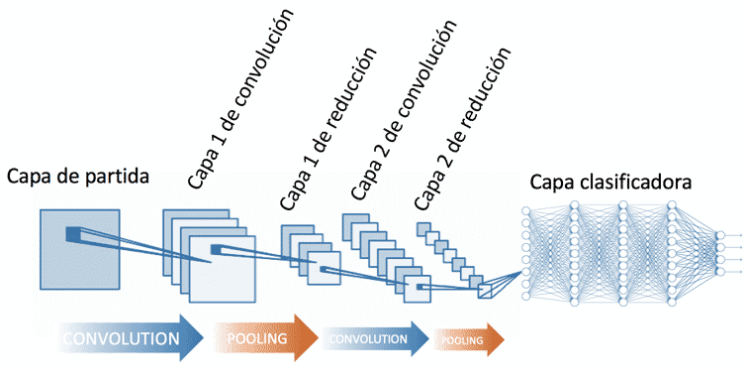

Figure 5: Red neuronal convolucional y sus diferentes capas.

Todas estas características convergen en un mismo objetivo que es la identificación de la voz humana. Los sistemas de reconocimiento de voz han ido evolucionando con el paso del tiempo y a medida de esto se han ido implementando diferentes algoritmos y métodos que han mejorado los resultados de la captación de señales. La base fundamental de todos estos sistemas parte de ASR.

7. Reconocimiento Automático del Habla (ASR)

Esta tecnología fue de las primeras en permitir la comunicación dentro una persona y una computadora convirtiendo las señales del habla a secuencias de texto. Un sistema ASR este compuesto por el proceso de las señales y extracción de características que lo que hace es recibir una señal de audio como entrada en primer lugar tratando de limpiar la señal de ruidos externos pasándola a una señal de dominio de frecuencia extrayendo las características mas relevantes. Después mediante el modelo acústico con conocimientos de fonética y acústica genera un modelo las características en secuencia que se envía a un modelo de lenguaje que realiza la estimación de probabilidad con secuencias de palabras hipotéticas para finalmente, mediante búsqueda de hipótesis, mezclar los dos puntos anteriores para generar el texto en base al aprendizaje, esto si dependiendo del nivel de dominio que tenga el modelo. Algo bastante simple y lo mas parecido a una red de aprendizaje, pero sin uso de técnicas de aprendizaje profundo. (Díaz, J. A., 2003)

El gran problema que sufría este modelo era su método de clasificación basado en métodos estadísticos que, al seguir con esta tendencia de uso, a lo largo del tiempo perdía robustez y existía inconvenientes dependiendo del entorno en el que se tomaba la muestra, además del locutor y demás parámetros externos. Estos problemas fueron abarcados como una capa de salida para redes neuronales recurrentes. Esta capa la llamaron Connectionist Temporal Classification que lo que hacia era clasificar toda esta información entrante basura como temporal, etiquetándola al no encontrar la alineación entre las etiquetas de entradas y las etiquetas de destino. (Díaz, J. A., 2003)

II. Metodología

La metodología propuesta en este articulo se compone de 5 fases en esencia: Data Gathering, Data Wrangling, diseño de red neuronal CNN y entrenamiento, predicciones y por último análisis y recomendaciones. Todas estas fases serán realizadas en el entorno de Python y aplicaciones nativas de iOS para grabación de audio.

1. Data Gathering (Recolección de datos)

Muchos otros sistemas de reconocimiento de voz se basan en los fonemas; el presentado aquí ignora por completo esta característica, centrándose solo en las frecuencias de cada una de las voces.

Para la extracción de la información se tomo una muestra pequeña de 6 individuos de diferente sexo y con diferente timbre de voz para poner a prueba la red neuronal. A cada uno de ellos se los puso en frente de un micrófono unidireccional SHURE PG58 para captar las ondas de la voz mediante una interfaz de audio FOCUSRITE 2i2 conectada a la Laptop. Mediante un programa llamado Logic Pro se grabaron las voces para exportarlos en formato WAV, escogiendo este formato ya que es uno de los formatos mas puros de audios al no realizar compresiones en su exportación. Se hizo de esta manera ya que mientras mas fiable y de mejor calidad es la fuente de información, mas precisa será la extracción de datos y por ende su aprendizaje y predicción.

Por cada individuo se tomo una muestra de 10 audios con diferentes frases y diferentes tiempos con un total de 60 audios. Las frases utilizadas fueron las siguientes:

- Buenos días
- El sabio crea los demás copian
- Si tiene prisa tropieza
- Habla menos observa mas
- Quiérete es gratis
- Se feliz no aceptes menos
- Trae tu propio sol
- Todo tiene su tiempo
- Escucha tu alma
- Cultiva tu amistad

Para poder identificar de mejor manera a nuestros individuos a lo largo del proyecto aclararemos sus rangos de audios con sus nombres respectivamente.

- DANIEL (audio 1 – audio 10)
- FABIS (audio 11 – audio 20)
- PEDRO (audio 31 – audio 40)
- TATI (audio 41 – audio 50)
- PAZ (audio 21 – audio 30)
- TEFA (audio 51 – audio 60)

Una ves exportada toda la data de audios procedemos a su tratamiento para convertirlos en señales digitales que puedan ser interpretadas y tratadas. Para eso utilizamos una librería para análisis de audio y música en Python, LIBROSA.

Para ello utilizamos la propiedad load de esta librería y mediante MATPLOTLIB y otra propiedad de la librería LIBROSA que nos muestra la señal de la voz gráficamente

DISPLAY.WAVEPLOT representamos una muestra de las voces de una misma frase, en este caso *"Cultiva tu amistad"*, para verificar que estas se encuentren cargadas. De esta manera podemos analizar las diferencias que existe entre ondas, siendo evidente la variación de sus frecuencias.

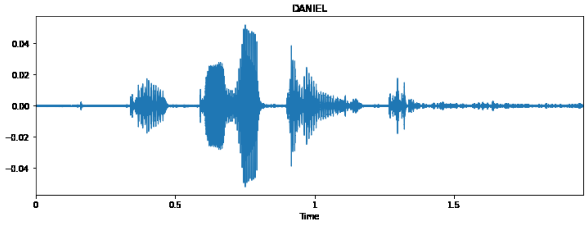

Figure 6: Espectro de voz Daniel

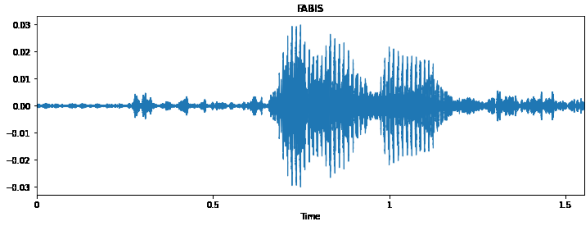

Figure 9: Espectro de voz Fabis

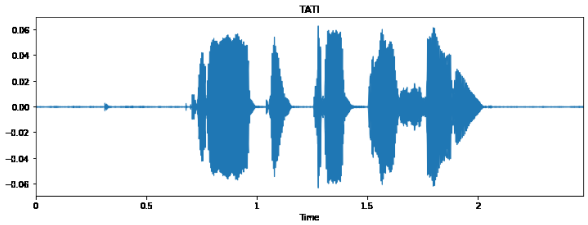

Figure 7: Espectro de voz Tati

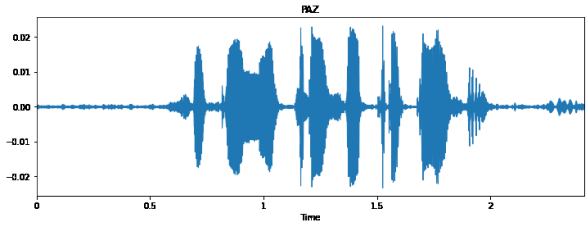

Figure 10: Espectro de voz Paz

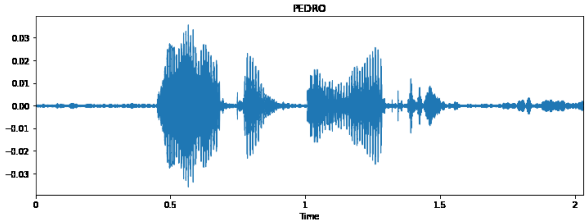

Figure 8: Espectro de voz Pedro

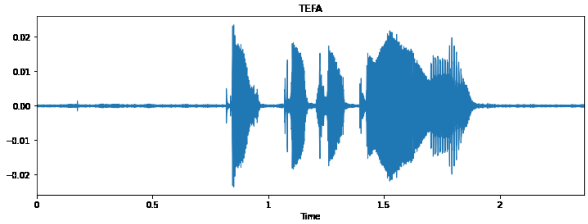

Figure 11: Espectro de voz Tefa

2. Data Wrangling (Tratamiento de datos)

Para preparar los datos haremos uso netamente de la librería LIBROSA a la par de NUMPY, de esta manera extraeremos los datos generados por los cálculos matemáticos y de los mismos el calculo de la media; generando desde esta sección un identificador por audio. De esta manera obtenemos los indicadores necesarios para la alimentación de la red Neuronal. De cada uno de los audios extraeremos los siguientes parámetros.

a. LIBROSA MFCCS

De esta forma, según la teoría del coeficiente de frecuencia ceptral de Mel, transformamos la información recibida del audio, a frecuencias similares a lo que un odio humano puede captar en los intervalos en Mels. Utilizamos un numero de coeficientes de 20; el valor que es utilizado de manera amplia en el reconocimiento del habla es 12, pero en este caso realizaremos un cambio al contar con mayor

cantidad de características extraídas por lo que necesitaremos mas información para evitar confusión y ambigüedad entre parámetros.

b. LIBROSA STFT o Transformada Corta de Fourier.

Generamos la transformada de Fourier, pero en este caso de tiempo corto por el hecho de encontrarnos trabajo con señales de frecuencia variables en el tiempo, para lo cual este tipo de transformada nos ayudó a generar los diferentes coeficientes las ondas senoidales presentes en el audio.

c. LIBROSA chroma_stft (Cromagrama de STFT)

Mediante esta función se calcula un cromagrama a partir de los datos obtenidos con STFT. Un cromagrama versus a un espectrograma común y corriente consiste en acumular los componentes de una frecuencia que pertenecen a una misma nota musical. De esta manera podemos definir presencia de armónicos y la frecuencia/nota dominante en cada voz.

d. LIBROSA MelSpectogram (Espectograma de Mel)

Mediante esta función simplemente transformamos toda la señal que se encuentra en la escala de Hz a escala de Mel.

e. LIBROSA spectral_contrast

Aplicamos el realce del espectro de la señal de frecuencia generada con la STFT con el objetivo de resaltar las frecuencias dominantes y filtrar aun más la información del audio.

f. LIBROSA tonnetz

Tonnetz nos sirve para extraer el centroide tonal de los armónicos de la señal, para eso aplicamos la detección de armónicos mediante una propiedad de librosa.effects.harmonic de la señal de audio para, mediante esos valores, obtener el tono central de la voz en cuestión.

g. LIBROSA pyin

Mediante esta función generamos la extracción de la frecuencia fundamental de la voz, el armónico fundamental, la frecuencia de la cual parte el resto de los armónicos y nos brinda el tono de voz. Vamos a desglosar esta propiedad para una mayor comprensión.

YIN es un método basado en la auto correlación para la estimación de la frecuencia fundamental. En primer lugar, se calcula una función de diferencia normalizada sobre cuadros de audio cortos (superpuestos). A continuación, se selecciona el primer mínimo en la función de diferencia por debajo de umbral como una estimación del período de la señal. Finalmente, el período estimado se refina mediante interpolación parabólica antes de convertirlo en la frecuencia correspondiente.

pYIN es una modificación del algoritmo YIN para la estimación de la frecuencia fundamental (F0). En el primer paso de pYIN, los candidatos a F0 y sus probabilidades se calculan utilizando el algoritmo YIN. En el segundo paso, se usa la decodificación de Viterbi para estimar la secuencia F0 más probable y los indicadores de voz.

El algoritmo de Viterbi selecciona la señal con un mayor valor de credibilidad, esto equivale a la menor distancia entre señal y señal. Esto se lo conoce como distancia Hamming. Cada una de las ramas creadas se etiquetan con la distancia, el algoritmo procesa la secuencia que recibe de manera iterativa, comparando los trayectos. Los que tienen un trayecto menor son los que sobreviven.

3. Diseño CNN y entrenamiento

Para la red neuronal primero dividimos todos los datos en dos grupos, **TRAIN con 56 audios** y **TEST con 12 audios**, 2 audios por cada una de las personas dentro del experimento. En total obtendremos 194 indicadores que serán ingresados a la red neuronal. Aplicamos un método de normalización para nuestras variables, para eso utilizamos la librería de skitlearn.

Vamos a agregar en este caso 3 capas con activación relu, el primer relu contará con una densidad 194 que son el total de indicadores que obtuvimos en la extracción de datos anterior y un dropout de 0.1. El segundo y tercero relu contará con una

densidad de 128 con un dropout de 0.25 y 0.5 respectivamente; y por ultimo la cuarta capa que contará con una función de activación softmax con una densidad del numero de indicadores únicos en el experimento, siendo el caso, 6 indicadores. Para su optimización aplicaremos el optimizado de learning Adam.

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense (Dense)                (None, 194)               37830

 dropout (Dropout)            (None, 194)               0

 dense_1 (Dense)              (None, 128)               24960

 dropout_1 (Dropout)          (None, 128)               0

 dense_2 (Dense)              (None, 128)               16512

 dropout_2 (Dropout)          (None, 128)               0

 dense_3 (Dense)              (None, 6)                 774

=================================================================
Total params: 80,076
Trainable params: 80,076
Non-trainable params: 0
_________________________________________________________________
```

Figure 12: Diseño de la Red Neuronal Convolucional

Dentro de su entrenamiento aplicamos diferentes combinaciones de valores de batch y epochs con el fin de encontrar la combinación ideal para conseguir un accuracy que pueda garantizar una predicción confiable. Las combinaciones se realizaron con los siguientes valores.

Batch	**Epochs**
256, 512, 1024	50, 100, 200, 300

4. Predicciones

Una ves que se obtuvo el accuracy mas alto con la combinación de los valores de batch y epochs, nos podemos fijar que la red neuronal predice perfectamente con un grado alto de asertividad. Dentro de estos existen dos fallos de reconocimientos dentro de una voz femenina y masculina respectivamente.

	id	sexo	persona	entorno	preds
0	audio9.wav	0	DANIEL	test	PEDRO
1	audio10.wav	0	DANIEL	test	DANIEL
2	audio19.wav	0	FABIS	test	FABIS
3	audio20.wav	0	FABIS	test	FABIS
4	audio29.wav	1	PAZ	test	PAZ
5	audio30.wav	1	PAZ	test	PAZ
6	audio39.wav	0	PEDRO	test	PEDRO
7	audio40.wav	0	PEDRO	test	PEDRO
8	audio49.wav	1	TATI	test	TATI
9	audio50.wav	1	TATI	test	TATI
10	audio59.wav	1	TEFA	test	TEFA
11	audio60.wav	1	TEFA	test	PAZ

Figure 13: Predicción de la red neuronal convolucional

III. Análisis de resultados y recomendaciones.

Como pudimos fijarnos en los resultados tanto los sujetos Fabis, Paz, Pedro y Tati fueron predichos exitosamente, pero existe una falla dentro de Daniel y Tefa. ¿Cuál es el motivo de este error?

Los sistemas de reconocimiento de voz tienen cierta sensibilidad y falla ante voces patológicas o la claridad de emisión de la voz de locutor. En este caso, tanto Pedro como Daniel, el momento de realizar las grabaciones, tendieron a bajar potencia vocal a comparación de su potencia de habla normal, estas actitudes se pueden presentar por diversos factores externos, como la presión de encontrarse frente a un micrófono, o el estrés o presión de emitir su voz lo mas natural posible, dando como resultado algo completamente contrario a lo esperado.

El ruido puede ser un factor esencial y perjudicial al momento de realizar el tratamiento de voz, en el caso especifico de Daniel, su tipo de voz es áspera y de poca potencia, pronunciando las palabras de manera que pueden pasar como ruido cuando no es así. Por otro lado Pedro tendía a cambiar la posición del micrófono, afectando de cierta manera la calidad del audio captado, además de encontrarse un poco ronco.

Con respecto a Tefa, el sujeto contaba con una voz con poca potencia. Previo a la toma de la muestra, intencionalmente se espero que el sujeto se encuentre en una situación donde

la voz se puede ver afectada, en este caso, al momento de despertar, además en ciertos audios se le solicito de manera intencional al sujeto que suba un poco la potencia vocal. Sus grabaciones cuentan con ciertas secciones ásperas, roncas, una voz patológicamente afectada lo cual afecta a la toma y extracción de características de manera clara.

Por otro lado, Tatiana, cuenta con una voz de gran potencia, además de claridad, pero ¿Por qué el sistema lo confunde con una voz completamente opuesta? Esto se puede dar por el mismo hecho de las características de su voz. Intencionalmente en ciertos audios pedí al individuo que moderé el volumen de su voz, siendo este un cambio bastante significativo con respecto a los otros audios que se tomaron con la potencia natural del sujeto, por lo que el sistema al ver variedad de potencia e inconscientemente, por parte del individuo, cambio de tono, el sistema confunde sus medias de características extraídas, dando como resultado una predicción errónea.

Como podemos observar las voces que fueron confundidas tienen ciertas características relacionadas en común, como el cambio de potencia o la patología que afectaba la voz en ese momento. Todo esto afecta a los resultados de la red neuronal.

IV. Conclusiones

En el desarrollo del presente articulo académico se presentaron varias dificultades, una de las mas grandes fue la arquitectura del chipset que presenta mi ordenador. El chip M1 no es un CPU como la arquitectura de un Intel, este es un SoC (System on a chip). El problema radicaba en que no muchas librerías o herramientas están adaptadas a la actualidad para este tipo de procesadores. Un claro ejemplo esta en las librerías de Python, las cuales pudieron ser de gran utilidad en mi desarrollo, como lo es PyAudio o WAVE, grandes librerías para tratamiento de audio en tiempo real o DeepSpeech una red de aprendizaje profundo en la cual no se necesita de un diccionario para su aprendizaje. La solución propuesta fue utilizar el entorno de Google Colab, un entorno muy completo de desarrollo en Python pero que en el cual, lastimosamente no funcionaban estas librerías de igual forma.

Posteriormente dentro de todo el campo referente a la investigación y tratamiento de voz como un sistema biométrico, existía carencia de información libre que pueda ser de utilidad a mas de tratar con CNN. Existían proyectos varios muy interesante pero la limitación radicaba en los pagos que se tenían que realizar para poder acceder a esa información. Aplicaciones con DNN, RNN; la mas llamativa desde mi perspectiva una investigación titulada *"DeepVoice: Tecnologías de Aprendizaje Profundo aplicadas al*

proceso de voz y audio" el cual es un proyecto que estaba centrado en investigar nuevas formas y algoritmos para el entrenamiento en arquitecturas de aprendizaje profundo (Martin Calle, 2019) algo muy interesante pero que lastimosamente no continuo con su desarrollo. De igual manera un estudio titulado *"End-to-end Deep Learning para reconocimiento automático del habla"* el cual aplicaba la librería DeepSpeech para la captación de señales en la cual entrenaba una RNN optimizada con GPU (A. Hannun 2014). Estos y muchos otros ejemplos se dieron en el campo, pero finalmente, de toda complicación, siempre se puede encontrar algo mejor.

La librería LIBROSA es un entorno completo para el tratamiento de señales de audio y música, lo cual facilitó de una manera muy favorecedora el desarrollo de mi investigación. Aprovechando los beneficios incorporados en la Macbook Pro por la potencia de captación de audio y con mi equipo de sonido, pude lograr desarrollar un sistema prototipo para el reconocimiento de la voz utilizando técnicas de aprendizaje profundo aprovechando el potencial de la librearía aplicada en el presente proyecto. El desarrollo pudo haber ido un poco mas allá con la opción de poder captar voces en tiempo real, pero me vi limitado tanto por tiempo como por el acceso limitado a librerías con las cuales no podía trabajar debido al entorno de Google Colab.

Algo a tomar en cuenta y con lo cual se puede llegar a trabajar en un futuro, es el filtrado correcto de ruido e información de frecuencias basura. Una aplicación muy interesante presente en la investigación titulada *"Algoritmo robusto para la detección de la frecuencia fundamental de la voz basado en el espectrograma"* (Díaz, J. 2003) propone un algoritmo que capta la frecuencia fundamental dentro de Matlab tratando de controlar la filtración de ruido. Esto depende mucho del entorno en el que se tomen las muestra y la tecnología que se utilice para ello.

Espero que todas estas limitaciones, específicamente las limitaciones de compatibilidad de librerías se solventen en un futuro, seria muy interesante poder aprovechar todo el potencial que un SoC puede ofrecer para en el ámbito de inteligencia artificial y aprendizaje profundo como lo es el SoC M1 Pro desarrollado por Apple.

Finalmente espero mi investigación y desarrollo sea de utilidad para futuros proyectos que se encuentren dentro del mismo objetivo investigativo o derivados.

V. Bibliografía

Barrios, K., López, J., Mendieta, S., Benavides, R., & Sáez, Y. (2018). Sistema de reconocimiento de voz: un enlace en la comunicación hombre-máquina. Revista De Iniciación Científica, 4, 92-95. https://doi.org/10.33412/rev-ric.v4.0.1827

Chan, Y. T., Lavoie, J. M. M., and Plan, J. B., A Parameter Estimation Approach to Estimation of Frequencies of Sinusoids, IEEE Transactions on Acoustics, Speech and Signal processing, Vol. ASSP-29 No. 2, p. 214-219, Abril, 1981.

Yannis, S., Decomposition of Speech Signals into a Periodic and Non-periodic Part Based on Sinusoidal Models, Proceedings of the IEEE International Conference on Electronics, Circuits, and Systems, Vol. 1, p. 514-517, 1996.

Díaz, J. A., Sapienza, C., Rothman, H. B., & Natour, Y. (2003). Algoritmo robusto para la detección de la frecuencia fundamental de la voz basado en el espectrograma. Revista Ingeniería UC, 10(3), 7-16.

Rufiner, H. L., & Milone, D. H. (2004). Sistema de reconocimiento automático del habla. Ciencia, Docencia y Tecnología, 15(28), 151-177.

Milone, D. H. (2005). Reconocimiento automático del habla con redes neuronales artificiales. Ciencia, Docencia y Tecnología, 16(31), 261-322.

A. Hannun et al.,(2014), Deep Speech: Scaling up end-to-end speech recognition.

Cruz, I. B., Martínez, S. S., Abed, A. R., Ábalo, R. G., & Lorenzo, M. M. G. (2007). Redes neuronales recurrentes para el análisis de secuencias. Revista Cubana de Ciencias Informáticas, 1(4), 48-57.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference (Vol. 8, pp. 18-25).

Ryan, M. S., & Nudd, G. R. (1993). The viterbi algorithm.

Hasan, M. R., Jamil, M., & Rahman, M. G. R. M. S. (2004). Speaker identification using mel frequency cepstral coefficients. variations, 1(4), 565-568.

Rincón, C. (2007). Diseño, implementación y evaluación de técnicas de identificación de emociones a través de la voz (Doctoral dissertation, Tesis de pregrado). Recuperado de http://lorien. die. upm. es/barra/pfcs/2007-carmenr/docs/proyecto. pdf)

Bernal, J., Gómez, P., & Bobadilla, J. (1999). Una visión práctica en el uso de la Transformada de Fourier como herramienta para el análisis espectral de la voz. Estudios de fonética experimental, 75-105.

Echeverry, J. D., Lemus, C. G., & Orozco, Á. Á. (2007). Análisis de la densidad espectral de potencia en registros mer. Scientia et technica, 1(35).

De Cheveigné, Alain, and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music." The Journal of the Acoustical Society of America 111.4 (2002): 1917-1930.

Mauch, Matthias, and Simon Dixon. "pYIN: A fundamental frequency estimator using probabilistic threshold distributions." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

Jaramillo, Á., & Varona, R. L. (2007). Transformada corta de Fourier. Scientia et technica, 1(34).

Benito Gorrón, D. D. (2017). Detección de música en contenidos multimedia mediante ritmo y armonía (Bachelor's thesis).

Yang, J., Luo, F. L., & Nehorai, A. (2003). Spectral contrast enhancement: Algorithms and comparisons. Speech Communication, 39(1-2), 33-46.

Tymoczko, D. (2012). The generalized tonnetz. Journal of Music Theory, 1-52.

Martín Calle, I. (2019). Reconocimiento de voz basado en características DNN Bottleneck (Master's thesis).