Análisis del Comportamiento de la Voz Humana para Detección de Depresión usando Redes Neuronales Convolucionales

Carlos Enmanuel Espinoza-Vicuña, Pregrado en Ciencia de la Computación[1], y Yuri Nuñez-Medrano MSc. en Ingeniería de Sistemas [2]
[1,2] Universidad Nacional de Ingeniería, Perú
[1]*cespinozav@uni.pe,* [2]*ynunezm@uni.edu.pe*

Abstract*– Este artículo muestra los resultados de una investigación realizada para el reconocimiento de depresión a través del análisis de voces grabadas en entrevistas psicológicas. Para lograr ello, se han usado técnicas de procesamiento de voz y modelos de inteligencia artificial. Los audios fueron obtenidos de la base de datos DAIC-WOZ. Uno de los primeros desafíos, fue el tratamiento de los audios. Esto debido a dos razones. En primer lugar, los audios tienen una duración considerable, de más de 20 minutos en muchos casos, lo que provoca una dificultad al poder describirlos. En segundo lugar, la mayoría de los audios tienen dos tipos de voces, que pertenecen al agente entrevistador y al participante. Por lo tanto, se emplearon métodos de segmentaciones, como la Diarización de voces o segmentaciones de características específicas. Ello con el fin principal de discriminar la voz del agente entrevistador y quedarse únicamente con la voz del participante. Luego de limpiar los audios, se observó que los audios aún eran extensos. Para ello, se recurrió a diferentes formas de extraer características relevantes en cada audio, transformándolos en espectrogramas que se ajustaron mejor al estudio. Finalmente, se usaron estas representaciones de audios como entrada en el modelo de red neuronal convolucional usado. Así mismo, para mejorar los resultados y reducir el overfitting, se emplearon técnicas como data augmentation. Durante ello, se revisaron recurrentemente los pasos previos de la metodología. Al final, se evaluó el modelo.

Keywords*— Depresión, Inteligencia Artificial, DAIC-WOZ, Segmentación, Diarización, Espectrogramas, Red Neuronal Convolucional, Overfitting, Data Augmentation.

I. Introducción

La depresión no es un problema reciente. Por lo contrario, ha estado siempre presente a lo largo de la historia humana. En el siglo IV a.c. Hipócrates comenzó a ver este tipo de problema al que llamó: 'Estado de ánimo pasajero'. En ese entonces, concluyó que se trataba de una enfermedad causada por desajustes en la bilis negra. Explica Hipócrates que el hombre es una unidad psicosomática y que todas las enfermedades son consecuencia de un desequilibrio humoral [3]. Según la OMS: "La depresión es distinta de las variaciones habituales del estado de ánimo y de las respuestas emocionales breves a los problemas de la vida cotidiana. Puede convertirse en un problema de salud serio, especialmente cuando es de larga duración e intensidad moderada a grave, y puede causar gran sufrimiento y alterar las actividades laborales, escolares y familiares. En el peor de los casos puede llevar al suicidio. Cada año se suicidan cerca de 800 000 personas. Siendo esta, la segunda causa de muerte en personas de 15 a 29 años" [4].

Es posible realizar una estimación del estado depresivo del paciente usando la escala de Hamilton. Este resultado, se puede lograr a través de los cuestionarios conocidos como PHQ-8 y PHQ-9 [20]. El interés por diagnosticar la depresión haciendo uso de nuevas herramientas tecnológicas es notorio. Ello, debido a una amplia gama de publicaciones referente a ello [20] [21] [22] y [23]. Estas investigaciones están enfocadas en analizar al habla deprimida. Se observó que los pacientes deprimidos demostraban consistentemente anomalías prosódicas en el habla, encontrando una variación reducida en volumen, inflexiones repetitivas de tono, patrones de acentuación, tono y volumen monótono. Cada año se realizan las competencias denominadas AVEC (Audio/Visual Emotion Challenge). Ahí, se evalúan proyectos relacionados al uso de inteligencia artificial analizando comportamientos en personas que padecen trastornos del estado de ánimo y desorden de estrés postraumático (PTSD) como la depresión. La base de datos Distress Analysis Interview Corpus (DAIC) es muy usada en estas competencias. Ya que brinda los datos propicios, especialmente audios y videos de entrevistas psicológicas, constantemente actualizadas [24]. Cada investigación, busca y emplea una metodología diferente. Dado que cada vez, existen novedosas formas de tratar y extraer las características necesarias de audios. Además, se ha probado con diferentes modelos usando redes neuronales convoluciones (CNN), redes neuronales concurrentes (RNN), modelo Gaussian Mixture (GMMs) y técnicas de aprendizaje como Transfer Learning. En el presente proyecto, se trató de encontrar una metodología donde se use nuevas herramientas de preprocesamiento de audios y modelos de inteligencia artificial, para obtener mejores resultados.

II. Objetivos

El desarrollo de la siguiente investigación está enfocado en los siguientes objetivos:

1. Aplicación de segmentación, con la finalidad de analizar los audios de la base de datos Daic-Woz. Especialmente, aplicar métodos de diarización usando servicios de Cloud Computing.
2. Uso técnicas de data augmentation en los audios para obtener una mayor cantidad de datos que serán usados

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2021.1.1.491
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

en el entrenamiento, validación y testeo del modelo de red neuronal usada. Esto, con la finalidad de reducir el overfitting.

3. Extracción espectrogramas que caracterizan partes esenciales de los audios.
4. Desarrollo y pruebas de redes neuronales convolucionales para la clasificación de estados de depresión.

III. Estado Del Arte

A. *Descripción de la Depresión.*

La depresión, es un trastorno que todas las personas han experimentado en algún momento de desequilibrio en el estado emocional. Según la Organización Panamericana de Salud (OPS): "La depresión, es la principal causa de problemas de salud y discapacidad en todo el mundo. Según las últimas estimaciones de la Organización Mundial de la Salud (OMS), más de 300 millones de personas viven con depresión, un incremento de más del 18% entre 2005 y 2015." [1]. En la Figura 3.1, se observa la distribución de casos depresivos a nivel mundial. La depresión tiene diferentes causales como factores genéticos, psicológicos, ambientales, biológicos, etc [2]. Sin el tratamiento y el control adecuado, este mal puede somatizar en otras enfermedades relacionadas al sistema nervioso como Parkinson y enfermedades carenciales como Hiper/hipotiroidismo. Además de otras enfermedades infecciosas tanto virales como bacterianas [6]. En el caso peruano, la situación es preocupante. Dado que como en la mayoría de los países subdesarrollados, los diagnósticos y tratamientos para la depresión aún son muy precarios. Además, según Estudios Epidemiológicos de Salud Mental (EESM) cada año en promedio un 20.7% de la población mayor a 12 años padece de algún tipo de trastorno mental. Dentro de esta población: "Los trastornos más frecuentes son los episodios depresivos, con una prevalencia anual que varía del 4% en Lima rural y 8,8% en Iquitos; y, se estima un promedio nacional de 7,6%." [7].

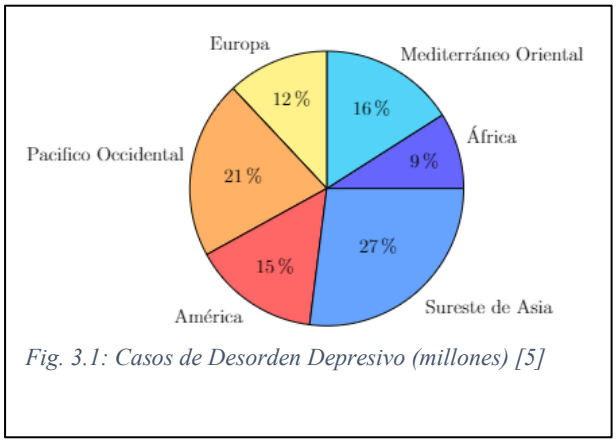

Fig. 3.1: Casos de Desorden Depresivo (millones) [5]

B. *Tipos de Depresión.*

a) Depresión Mayor: Los episodios depresivos mayores pueden manifestarse en diferentes etapas de la vida y de forma espontánea. Afecta hasta el 10% de los jóvenes. De los cuales, el 90% de los casos son jóvenes. Por lo general, estos pacientes se recuperan del episodio depresivo en 1 a 2 años. Sin embargo, los episodios pueden reaparecer en el 40% - 70% de los casos [10]. Algunos síntomas de este tipo de depresión son: persistente estado de ánimo triste la mayor parte del día, pérdida de interés o placer en pasatiempos y actividad, sentimientos de culpa, inutilidad, impotencia, disminución de energía, fatiga, dificultad para concentrarse, recordar o hacer decisiones, inquietud e irritabilidad [12].

b) Trastorno Depresivo Persistente (PDD): Puede darse por largas etapas, aunque es menos frecuente que la depresión mayor. Implica los mismos síntomas; estado de ánimo triste combinado con poca energía, falta de apetito o comer en exceso, e insomnio o sueño excesivo. Puede aparecer como estrés, irritabilidad y anhedonia leve (incapacidad de obtener placer de la mayoría de las actividades) [11].

c) Trastorno Bipolar: Conocido también como depresión maníaca. Se caracteriza por estados de ánimo que cambian de agudos a graves (manía) o agudos leves (hipomanía) a graves (depresión). Los episodios de ánimo asociados con el trastorno persisten de días a semanas o más y pueden ser dramáticos. Las personas con trastorno bipolar presentan comportamientos que varían dependiendo del estado ánimo [12].

C. *El Habla Deprimido e Inteligencia Artificial.*

La actividad facial, la prosodia del habla, los gestos, los movimientos de la cabeza y la expresividad son señales de comportamientos que están fuertemente ligados con la depresión [9]. Dado que este mal mental influye en como una persona se siente, piensa y coordina las actividades diarias como: dormir, comer, trabajar, escribir, hablar, entre otros [8].

Actualmente, hay mucho interés dentro del área de Inteligencia Artificial por analizar estas señales del comportamiento humano. "Las investigaciones paralingüísticas tempranas sobre el habla deprimido encontraron que los pacientes demostraron de manera constante anomalías del habla prosódica"[9]. Estas anomalías en el habla depresivo son:

a. Tono reducido.
b. Rango de tono reducido.
c. Frecuencia de habla más lenta.
d. Errores de articulación más altos [13].

D. *Entrevistas DAIC-WOZ.*

Distress Analysis Interview Corpus (DAIC), es una base de datos que colecciona información de una serie de entrevistas clínicas psicológicas. Estos datos acumulados, son usados en diferentes investigaciones para el diagnóstico de trastornos mentales como la depresión y ansiedad [17]. Antes de las entrevistas, los participantes rellenan un formulario de consentimiento de que sus datos recopilados serán usados con fines de investigación. La entrevista se encuentra totalmente diseñada por expertos para crear un ambiente de confianza. Al introducirlos en las sesiones, los participantes son estimulados

con una serie de imágenes y videos emocionales. Luego, los participantes y los agentes entrevistadores interactúan. Las entrevistas son grabadas y durante ello, se recopila una serie de indicadores verbales y no-verbales que sirven como parámetros para la medición de trastornos mentales. Estas entrevistas se desarrollan de forma personal, son echas en modalidad frente a frente, teleconferencia, Wizard-of-Oz y autómata [17].

El cuestionario usado es llamado PHQ-8. Empieza primero con una serie de preguntas para tomar confianza con el participante. Luego, siguen preguntas más personales con polaridad de valencia variable que consta de una fase positiva, otra negativa y una neutral. La fase positiva consta de preguntas como: "¿Cuáles dirías que son algunas de tus mejores cualidades?" o "¿Cuáles son algunas cosas que generalmente te ponen de buen humor?". La fase negativa consta de preguntas como: "¿Tienes pensamientos perturbadores?" o "¿Cuáles son algunas de las cosas que realmente te enojan?". Las preguntas neutrales incluyen: "¿Cuántos años tenía cuando se enlistó?" o "¿Qué estudiaste en la escuela? " [16].

El rango de resultado del cuestionario PHQ-8 está entre 0 y 24. El resultado binario de PHQ es 0 y 1. En ambos casos, valores mínimos corresponden a un estado bajo de depresión y los valores altos corresponden a un estado alto de depresión.

E. Segmentación de Audios.

La segmentación, es un proceso que consiste en separar una señal de audio digital en ciertos segmentos. Cada uno con características distintas como el habla, la música, los sonidos no verbales de la actividad humana, sonidos de animales, los sonidos ambientales, ruidos, silencio, etc [18] y [19]. La arquitectura general de la segmentación de audio consta de los siguientes pasos básicos:

a) Extracción de características: Para comenzar, la entrada de audio se corta inicialmente en cuadros superpuestos de muestras de audio y para cada cuadro se extrae un vector de características paramétricas. La secuencia calculada de vectores de características se reenvía a un módulo de detección inicial.

b) Detección inicial (etapa opcional): Se interpola a la estructura principal por dos razones. La primera razón es eliminar las partes de silencio antes de la etapa de segmentación, en lugar de utilizar una clase de "silencio". La segunda razón es descartar las partes de la señal que están fuera de interés (por ejemplo, en la tarea de segmentación del hablante solo se necesitan las partes del discurso para la etapa de segmentación). Por lo general, la detección del silencio y el ruido respiratorio se realiza a partir de un detector simple basado en energía (VAD) y la detección de la música y el ruido se logra utilizando modelos de mezcla gaussianos (GMM).

c) Segmentación: Aquí, se segmentan las sub-secuencias (fragmentos) con características acústicas comunes. Para la etapa de segmentación se siguen dos enfoques principales llamadas técnicas basadas en la distancia y las técnicas basadas en modelos.

d) Post-procesamiento del segmento o suavizado (etapa opcional): Para refinar o suavizar los resultados de la segmentación automática. Esta etapa corrige los errores relacionados con los segmentos detectados con una duración menor que los umbrales definidos empíricamente [18].

Para separar o discriminar voces en audio con distintos de hablantes, se utiliza la Diarización. Para ello, se puede usar un api llamado Speech-to-Text. implementándolo en diferentes arquitecturas, acorde al uso al que se quiere dar. En la Figura. 3.2, se muestra la arquitectura con la que se procesaron los audios. Dentro de las configuraciones en el uso del servicio, se puede seleccionar el tipo de modelo de aprendizaje con que se quiera trabajar. Existen modelos para transcribir videos, llamadas telefónicas y automatic speech recognition (ASR) [26].

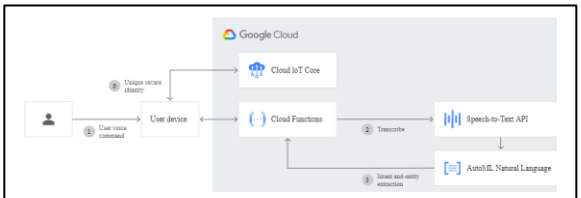

Fig. 3.2: Arquitectura para API Speech to Text, incorporando comando de voz [26].

F. Caracterización de Audios.

Para analizar los audios, se requiere obtener ciertos gráficos representativos que permitan describir en gran medida esos datos. Estos gráficos se pueden organizar en diferentes niveles de caracterización:

e) Forma de Onda (Time Domain Waveform): Este gráfico se obtiene haciendo uso de la Transformada de Fourier. La Transformada de Fourier (TF) hace referencia que toda función periódica de frecuencia w_0 puede expresarse como la suma infinita de funciones senos o cosenos de múltiplos enteros de n y w_0. Para la ecuación básica (1), se denomina w_0 a la frecuencia fundamental y a cada término seno o coseno se le conoce como armónica. Para la ecuación compleja (2), se descompone una señal periódica en senos y cosenos de diferentes frecuencias y amplitudes a través de la exponencial de Euler. Entre las formas de Fourier tenemos [14]:

FORMA BÁSICA: (1)

$$f(t) = a_0 + \sum_{n=1}^{\infty}[a_n cos(n\, w_0 t) + sen(n\, w_0 t)\,]$$

FORMA COMPLEJA: (2)

$$f(t) = a_0 + \sum_{n=1}^{\infty}[c_n e^{n\, w_0 t} + c_{-n} e^{-n\, w_0 t}\,]$$

$$f(t) = \sum_{-\infty}^{\infty}[c_n e^{(n\, w_0 t) j}]$$

dt

f) Si f (t) es una función para todo t real, integrable en el intervalo [t0 − T/2, t0+T/2]. Entonces, se puede obtener el desarrollo en serie de Fourier de f en ese intervalo. Fuera del intervalo, la serie es periódica con período T. Si f (t) es periódica para todo t real, la aproximación por series de Fourier también será válida en todos los valores de t.

g) Espectrograma: Permite representar señales en imágenes de diferentes tipos de frecuencias existentes. También, muestra la variación de niveles de energía con respecto del tiempo. El espectrograma de una señal de entrada se puede describir como el cuadrado de la magnitud de la Transformada de Fourier de corto tiempo (STFT) [25]. Cuya fórmula es como sigue:

ECUACIÓN: (3)

$$f(n, w) = \sum_{n=-\infty}^{\infty} x(i)w(n-1)e^{-j\omega n}$$

Donde:

1. x(i): Señal de entrada.
2. w(i): Momento en que n es una función tipo Hanning

h) Espectrograma MFCC: Esta caracterización está basada en Coeficientes Cepstrales de Frecuencia Mel. Este método es usado en el reconocimiento de voz basado en el dominio de frecuencia utilizando la escala Mel. Esta, es una escala musical perceptual del tono a juicio de observadores equiespaciados. Los MFCC usan un banco de kernels a escala Mel. Los kernels con frecuencia más alta tienen mayor ancho de banda que los kernels de frecuencia más baja. Pero, sus resoluciones temporales son las mismas. El último paso es calcular la Transformación discreta de coseno (DCT) de las salidas del kernel [15].

IV. Metodología De Investigación

A. Base de datos DAIC-WOZ.

Para tener acceso a la base de datos, se tuvo que firmar un acuerdo de uso EULA (End User License Agreement) y enviar al correo que se registra en su página web, perteneciente a University of Southerm California. Una vez enviada y aceptada la solicitud, se recibe la base de datos compartida. Esta base de datos está compuesta por:

a) Sección de etiquetas (labels): Contiene archivos que describen la colección de datos. Las entrevistas echas con el cuestionario PHQ8, genera resultados con el mismo nombre. Obteniendo así, la calificación dada de depresión. Estos archivos se llaman dev split, train split y test split. Todos contienen datos que describen las entrevistas realizadas a los participantes. Cuyos campos son: Participant_ID, Gender, PHQ_Binary, PHQ_Score, PCL_C (PTSD) y PTSD Severy.

b) Sección de Datos (data): Contiene 416 folders con nombres de la forma XXX _P, donde XXX es la ID de los participantes. Cada folder, contiene información detallada sobre la sesión realizada a un determinado participante. Encontrando, un folder llamado features, un script XXX_Transcipt.csv y el audio XXX_AUDIO.wav.

B. Preprocesamiento de Audios.

Para obtener los datos de entrada del modelo de red neuronal convolucional, se tiene que pasar por varias etapas. Durante estas, es importante tener en cuenta la estructura desarrollada en la descripción de la base de datos. A continuación, se describe las diferentes etapas.

c) Extracción de Datos: Cada folder de forma XXX_P, es descomprimida bajo la misma estructura de la base datos. Permitiendo diferenciar el conjunto de datos en data de entrenamiento (train_data), data de validación (dev_data) y data de testeo (test_data).

d) Diarización de audios: La diarización de audios, es otra forma más elaborada para discriminar las voces de los participantes y los agentes entrevistadores. Para ello se ha usado un api de google cloud llamado Speech-to-Text. El modelo predeterminado de ese servicio se ajustó bien al caso de uso que requiere la investigación. El primer inconveniente del api es que, al tener almacenados audios de forma local o en drive, sólo acepta solicitudes con duración de máximo de 1 minuto. Para solucionar esto, se ha creado un segmento de Storage. Haciendo uso de las instrucciones propias de este recurso, se copiaron ahí los audios para un mejor uso del API Speech-to-Text. Una vez almacenada los audios en Google Storage, es posible diarizar audios extensos. El propósito de usar Speech-to-Text, es obtener intervalos de tiempos de cada palabra mencionada en las grabaciones, la palabra y el tag (identificador que hace referencia a la persona emisora de la palabra). Es posible usar el api para extraer frases u oraciones. Pero, resultó más conveniente hacer una descripción del audio por palabras. Ya que así, se evita otros procesos de limpieza de ruido o silencio. La configuración usada es la siguiente:

```
client = speech.SpeechClient()
audio = types.RecognitionAudio(uri = gcs_uri)
config = speech.types.RecognitionConfig(
language_code='en-US',
enable_word_time_offsets=True,
enable_speaker_diarization=True,
audio_channel_count=1,
diarization_speaker_count=2)
operation = client.long_running_recognize(config, audio)
```

Se accede al audio, desde el segmento de Google Cloud usando la dirección **gcs_uri**, Se usa el modelo de reconocimiento de voz estándar. Se configura el idioma de voz a procesar en el audio **en-US**. Luego, se

habilita **enable_word_time_offsets** para obtener marcas de tiempo por palabra y el **enable_speaker_diarization** para habilitar la diarización. También, se define el número de canales usado para procesar el audio. Finalmente, se coloca el número **diarization_speaker_count** que corresponde a la cantidad de **tags** o hablantes en el audio. Teniendo en cuenta la duración y la cantidad de audios que se tuvo que procesar, el api se ejecutó durante dos días aproximadamente. En la Figura. 4.1, se observa el comportamiento de uso del API Speech-to-Text durante las solitudes de procesamiento. Cada solicitud, genera como resultado un archivo csv para cada audio, llamado XXX_Words_diarizations.csv. Este archivo consta del campo Text que almacena la palabra mencionada. Los campos Star_Time y End_Time que definen el intervalo de tiempo en que es mencionada cada palabra. Además, hay un campo Id_speaker, que es un identificador de cada hablante en un audio.

Fig. 4.1: Tráfico de datos durante uso de Speech-to-Text API.

a) Segmentación en Base a Audios Diarizados: Haciendo uso de los archivos dev Split, train Split y test Split, se ubican los audios de baja y alta depresión. Dado que a cada audio XXX_P, le corresponde un archivo XXX_Words_diarizations.csv. Se arma una lista de palabras para cada estado de depresión. Se realizan las búsquedas de las palabras en los archivos diarizados y se coteja si fue dicha por el participante. Luego, se hacen segmentaciones de cortes de 2 segundos. Para el caso de depresión alta, se usan palabras como: Depressed, Crime, Hate, Dead, etc.

b) Data Augmentation de Audios: Se realiza esta técnica con la finalidad de obtener más datos que sirvan como entrada para el modelo de red convolucional usada. Las características de estos datos están en base a los audios ya segmentados. Para estados altos y bajos de depresión, se hacen desplazamientos en los audios. Otra técnica importante es la reducción y el aumento de velocidad en los audios. Para el caso de depresión baja, se aumenta la velocidad y depresión alta se reduce. Esto siguiendo la lógica que las personas con depresión alta tienden a tener tonos más lentos y reducidos.

c) Extracción de Características de Audios: Haciendo uso de los scripts train_split.csv, dev_split.csv y test_plit.csv, se generan espectrogramas para data de entrenamiento, validación y testeo, respectivamente. Para ello, se usa la librería librosa, que ofrece el soporte para el análisis de audio. Se sigue el siguiente pseudocódigo:

```
#Get Spectrogram
get_spectrogram(wav):
    D = librosa.stft(wav, n_fft=480, hop_length=160,
    win_length=480, window='hamming')
    spect, phase = librosa.magphase(D)
    return spect

# Load audio and transform to spectrogram
y,sr=librosa.load(audiopath)
log_spect=get_spectrogram(y)
save_spectrogram(log_spect)
```

Se obtienen imágenes de espectrogramas de la siguiente forma (Ver Figura. 4.2 y Figura. 4.3):

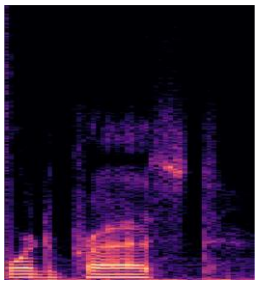

Fig. 4.2: Imágen de espectrograma

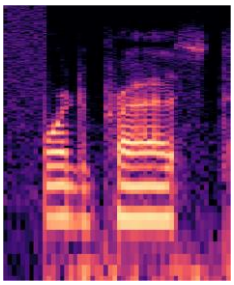

Fig. 4.3: Imágen de espectrograma a escala logarítmica.

C. Construcción de Modelo de Red Neuronal Convolucional.

El modelo creado consta de cuatro etapas convolucionales y una capa completamente conectada. Estas se describen a continuación:

1. Primera etapa convolucional: Es la capa de entrada. Cuya dimensión de entrada depende de la forma de los espectrogramas que se quiere entrenar. Luego, se agrega una capa convolucional con 88 kernels de 3x3. Inicialmente, se tuvieron 2 tipos de datos de entradas. La primera son imágenes de espectrogramas que pueden ser sin escala y con escala logarítmica de la forma (60,60,3). La segunda, son matrices stft bidimensionales de 513x52.
2. Segunda etapa convolucional: En esta etapa se realizan dos convoluciones seguidas. Cada una con 88 kernels de dimensión 3x3. Seguidas, de una función de activación relu. Luego, Se agrega una capa Add(). Esta, toma como entrada el resultado de la segunda convolución de esta etapa y el resultado de la convolución de la etapa anterior. Todos poseen la misma forma, y devuelve un solo tensor (también de la misma forma). Luego un sub-sampling maxpooling con kernel de 3x3 y un dropout con una tasa de 20%.

Lo que significa que una de cada 5 entradas se excluirá aleatoriamente en cada ciclo de actualización.

3. Tercera y Cuarta Etapa: De igual forma que la segunda etapa. La tercera y cuarta etapa, poseen las dos similares capas convolucionales, una capa Add() y una capa sub-sampling. Al final, para entradas de imágenes de espectrogramas de tamaño 60, queda una matriz resultante de dimensión 13x13.
4. Capa completamente conectada: Usando Flatten, se transforma un vector unidimensional de tamaño 14872. Luego este vector pasa por tres capas densas. La salida de la capa Densa se verá afectada por el número de neuronas (unidades) especificadas en la capa Densa. Por ejemplo, para la primera capa posee una salida de 256, la segunda de 128 y al final una salida con dos unidades. En cada capa se aplica BatchNormalization, para estandarizar las entradas y un Dropout() para mejorar los resultados del aprendizajes. Se utilizaron tasas de Dropout() entre el 20% al 50% de las neuronas. Una probabilidad demasiado baja tiene un efecto mínimo y un valor demasiado alto da como resultado un bajo aprendizaje por parte de la red. Se obtiene como salidas [1,0] para estado bajo de depresión (estado 0) y [0,1] para un estado alto de depresión (estado 1).

V. Análisis de Resultados

Al realizar varias pruebas con los tipos de entradas obtenidos (matrices stft e imágenes de espectrogramas). Se observó que el modelo de red convolucional tiene mejor comportamiento con imágenes de espectrogramas a escala logarítmica. A continuación, se muestran los resultados obtenidos. Este tipo de clasificación está diseñada para poder clasificar 2 clases. La clase 0 y la clase 1, corresponde a un estado bajo y alto de depresión, respectivamente (Llamados estado 0 y estado 1). Se tienen 1200 imágenes como data de entrenamiento, 250 para data de validación y 200 para data de testeo. En cada caso, siempre hay la mitad de los espectrogramas con etiqueta 0 y otra mitad con etiquetas 1. Las matrices son leídas de las imágenes de espectrogramas a escala logarítmica. La dimensión y forma de la entrada para el modelo es 60x60, con 3 canales. Antes del entrenamiento, matrices son permutadas con sus respectivas etiquetas en el mismo orden de permutación. Durante el proceso de entrenamiento, se guardan los mejores pesos y se detiene el proceso cuando no se observan mejoras. En la Figura. 5.1, se observa que realizan más de 60 épocas. El comportamiento de la gráfica es irregular, pero llega hasta 0.4 de pérdida (loss). De igual manera, en la Figura. 5.2, se observa un comportamiento de overfitting. El accuracy, en este caso llega hasta 0.8. El resultado del accuracy sobre los datos de testeo es de 0.78.

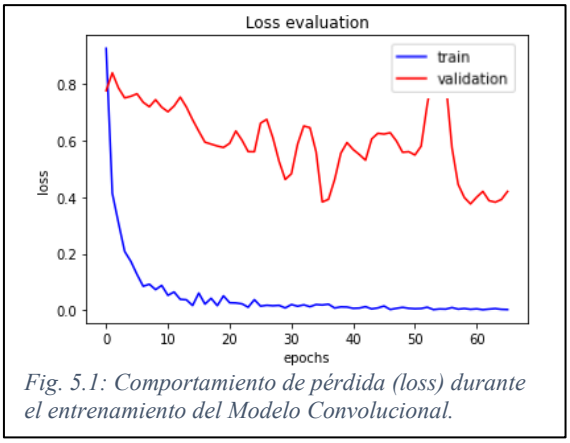

Fig. 5.1: Comportamiento de pérdida (loss) durante el entrenamiento del Modelo Convolucional.

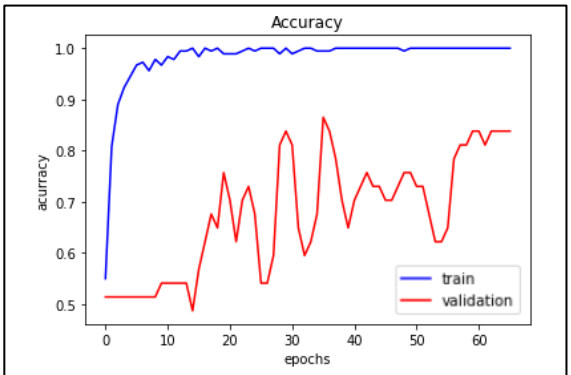

Fig. 5.2: Comportamiento de accuracy durante el entrenamiento del Modelo Convolucional.

VI. Discusión de Resultados

En general, los resultados obtenidos no fueron los esperados, pero bastante alentadores. Se observó un comportamiento de overfitting en las gráficas. Esto debido a dos razones. Primero, debido a que es necesario obtener más espectrogramas para cada clase de depresión, es posible mejorar empleando más técnicas de data augmentation para los audios. La segunda razón, es porque en realidad se debe asegurar que los espectrogramas deben tener características diferenciables entre cada clase. Estas características deben seguir una misma distribución en cada clase de depresión. Los espectrogramas a escala logarítmica han demostrado ofrecer mejores resultados. Usando espectrogramas a escala logarítmica, el accuracy para data de testeo es de 0.78. En comparación al accuracy para espectrogramas sin escala que resultó ser menor, 0.73. El proceso de preprocesamiento de voz al usar técnicas de diarización demostraron ser prometedores para extraer las características que se requerían. A pesar, de que en proceso hubo varios detalles a resolver como la duración considerable que tuvieron cada uno de estos audios que hacían difícil su análisis.

VII. Conclusiones

Se ha logrado implementar diferentes técnicas para el pre-procesamiento de los audios de la base de datos Daic-Woz. Por un lado, la diarización ha permitido describir a detalle cada entrevista grabada en los audios. Por otro lado, la segmentación de audio por tiempos permitió recolectar fragmentos específicos e importantes de audios. Luego, la aplicación de data augmentation modificando la velocidad de los audios acorde a la clase que pertenece cada audio, se acopla perfectamente a la característica de tono lento y reducido que poseen personas con depresión. El modelo de red convolucional usado ha mostrado tener gran potencial. En base a la metodología desarrollada en este proyecto. Se ha observado que se tuvo similar experiencia a otros trabajos realizados anteriormente. Debido que también experimentaron similares problemas con el overfitting. Para mejorar los resultados obtenidos, se puede investigar más acerca de modelos que permitan hacer una mejor clasificación. Recurrir a redes neuronales recurrentes, modelos Generative Adversarial Networks (GAN's) para generar espectrogramas de similares características, aplicar Transfer Learning o el uso de métodos bayesianos. También, es posible realizar procesamiento de texto usando Procesamiento de Lenguaje Natural (NLP).

VII. Agradecimiento

Se agrade a la Universidad Nacional de Ingeniería (UNI). Especialmente a la Facultad de Ciencias y al Vicerrectorado de Investigación de la UNI.

References

[1] Organización Panamericana de Salud. https://www.paho.org/hq/index.php?option=com_content&view=article&id=13102:depression-lets-talk-says-who-as-depression-tops-list-of-auses-of-ill-health&Itemid=1926&lang=es

[2] Julisca Cesar, Fáraz Chavoushi. “A Public Health Approach to Innovation”. Background Paper 6.15 Depression, Update on 2004

[3] J. L. González de Rivera, " Evolución histórica de la Psiquiatría". Psiquis, 1998; 19 (5):183-200.

[4] Organización mundial de la Salud (OMS), Detalles de Depresión, Recuperado de https://www.who.int/es/news-room/fact-sheets/detail/depression

[5] “Depression and Other Common Mental Disorders: Global Health Estimates”. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.

[6] S. Rivera Peñaranda, "Depresión y enfermedades somáticas". SEMERGEN. 2009;35 Supl 1:26-30

[7] Ministerio de Salud (MINSA), "Lineamientos de política sectorial en salud mental", Lima: Ministerio de Salud; 2018. 54 p.

[8] NIH (National Institute of Mental Health), Depresión, Información Básica, https://www.nimh.nih.gov/health/publications/espanol/depresion-sp/index.shtml

[9] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, Maja Pantic. "State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition”. AVEC’19, October 2019, Nice, France

[10] M Kovacs, T L Feinberg, M A Crouse-Novak, S L Paulauskas, R Finkelstein. "Depressive disorders in childhood. I. A longitudinal prospective study of characteristics and recovery". PMID: 6367688 DOI: 10.1001/archpsyc.1984.01790140019002.

[11] Harvard Health Publishing, Harvard Medical School, "Persistent Depressive Disorder (Dysthymia)", https://www.health.harvard.edu/a_to_z/dysthymia-a-to-z

[12] Anxiety and Depression Association of America, "Depression". url: https://adaa.org/understanding-anxiety/depression.

[13] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. "A review of depression and suicide risk assessment using speech analysis. Speech Communication". Volume 71, July 2015, Pages 10-49

[14] Jimy Alexander Cortes Osorio1, Andrew M. Knott2, José Andrés Chaves Osorio3, Departamento de Física, Universidad Tecnológica de Pereira, Pereira, Colombia, "Aproximación a la síntesis de la música". Scientia Et Technica; ISSN: 0122-1701.

[15] Roman A. Solovyev, Maxim Vakhrushev, Alexander Radionov, Vladimir Aliev, Alexey A. Shvets,"Deep Learning Approaches for Understanding Simple Speech Commands". October 2018. Available via license: CC BY 4.0

[16] Stefan Scherer, Member, IEEE, Gale Lucas, Jonathan Gratch, Member, IEEE, Albert Rizzo Member, IEEE, and Louis-Philippe Morency, Member, IEEE "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews". Published in *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59-73, 1 Jan.-March 2016, doi: 10.1109/TAFFC.2015.2440264.

[17] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, Louis-Philippe Morency. “The Distress Analysis Interview Corpus of human and computer interviews”. USC Institute for Creative Technologies, 12015 Waterfront Drive, Playa Vista CA 90094-2536, USA.

[18] I.J. Information Technology and Computer Science, 2014, 11, 1-9 "An Overview of Automatic Audio Segmentation". Published Online October 2014 in MECS.

[19] Aidan Hogg, Christine Evers and Patrick Naylor Electrical and Electronic Engineering, Imperial College London, UK. "Speaker change detection using fundamental frequency with application to multi-talker segmentation", May 16, 2019.

[20] Lang He, Cui Cao, "Automated depression analysis using convolutional neural networks from speech". Journal of Biomedical Informatics. Volume 83, July 2018, Pages 103-111.

[21] Himani Negi, Tanish Bhola , Manu S Pillai , Deepika Kumar , "A Novel Approach for Depression Detection using Audio Sentiment Analysis". Proceedings of 4th International Conference on Computers & Management (ICCM)2018.

[22] ZHIYONG WANG, LONGXI CHEN, LIFENG WANG, AND GUANGQIANG DIAO, "Recognition of Audio Depression Based on Convolutional Neural Network and Generative Antagonism Network Model" published in *IEEE Access*, vol. 8, pp. 101181-101191, 2020, doi: 10.1109/ACCESS.2020.2998532.

[23] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization". INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011.

[24] Michel F Valstar profile imageMichel Valstar, Björn W Schuller profile imageBjörn Schuller, Kirsty L Smith profile imageKirsty Smith, Timur R Almaev profile imageTimur Almaev, Florian Eyben profile imageFlorian Eyben, Jarek Krajewski profile imageJarek Krajewski, Roddy Cowie profile imageRoddy Cowie, Maja Pantic profile imageMaja Pantic, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge". AVEC '14: Proceedings of the 4th International Workshop on Audio/Visual Emotion ChallengeNovember 2014.

[25] Fatih Demir Firat University · Department of Electrical and Electronic Engineering 7.33, Daban Abdullah, Abdulkadir Sengur Firat University · Department of Electrical and Electronic Engineering, "A New Deep CNN model for Environmental Sound Classification". Published in *IEEE Access*, vol. 8, pp. 66529-66537, 2020, doi: 10.1109/ACCESS.2020.2984903.

[26] Documentación Speech to text., " Guía para inicio practico y conceptos ". url: https://cloud.google.com/speech-to-text/docs.