Reconocimiento de comandos de voz mediante espectrogramas y aprendizaje profundo

Pamela Salazar Espinoza
Estudiante Maestria Electronica
ITCR

***Abstract*—En el presente proyecto se propone un prototipo de reconocimiento de voz mediante el entrenamiento de redes neuronales.**

Keywords—espectrogramas, redes neuronales, aprendizaje profundo

I. Introduccion

Muchas caracterizaciones de las señales han sido propuestas en el pasado. Estas incluso han tenido altos grados de éxito, pero tienen un inconveniente; el proceso de selección manual de variables para cada paciente requiere de mucho tiempo y recursos. Esto ha llevado al planteamiento de algoritmos predictivos basados en redes neuronales. A cambio de un costo computacional elevado, las redes neuronales son capaces de determinar cuáles características de las señales son las más relevantes. [1] El espectro de poder de una serie de tiempo describe la distribución de poder en las diferentes frecuencias que la componen.

Existen diferentes algoritmos de extracción del espectro de poder de una señal, como las transformadas de Wavelet y de Fourier. A continuación, se describe la transformada de tiempo corto, una técnica que permite describir el comportamiento del espectro de una señal a lo largo del tiempo. [1]

Uno de los problemas dentro del reconocimiento de hablantes es la modificación de tono en la voz. En algunos casos, la persona a reconocer no puede controlar algunos cambios en su propia voz, por ejemplo, cuando una persona está enferma. Existen dos tipos de cambios en la voz: voluntarios e involuntarios. El primero ocurre cuando una persona, de manera consciente, hace alteraciones a su propia voz para no ser reconocido, e.g. hablar más grave, cubrir la boca o tapar su nariz. Los cambios involuntarios se dan cuando una persona no puede controlarlos dichos cambios, por ejemplo, cuando tiene un resfriado, tos o está ronco. Pueden haber algunas razones más para cambios involuntarios, sin embargo, estas son las más comunes. Puede haber otra clasificación para estos cambios, los naturales y los artificiales. Los cambios artificiales en la voz ocurren cuando, además del hecho de querer cambiar la voz, alguien utiliza un dispositivo para realizar este cambio, como un procesador de voz. Los Coeficientes Cepstrales en la Frecuencia de Mel, son los más usados para la parametrización de voz debido a su bajo costo computacional y a su robustez.[2]

II. La transformada de Fourier de tiempo corto

La transformada de Fourier consiste en deslizar una ventana temporal pequeña a lo largo del dominio temporal de una señal, calculando la transformada de Fourier en cada uno de los segmentos. El resultado es un conjunto de funciones de dominio de frecuencia, indexadas por el tiempo en el cual estaba centrada la ventana corta en la cual cada una fue calculada.[1]

III. Espectrograma

El espectrograma consiste en la representación gráfica del espectro de frecuencias de la emisión sonora. El espectrograma puede revelar rasgos, como altas frecuencias o modulaciones de amplitud, que no pueden apreciarse incluso aunque estén dentro de los límites de frecuencia del oído humano.[2]

IV. Escalas de Mel

Los Coeficientes Cepstrales en la Escala de Mel (MFCC) representan la amplitud del espectro del habla de manera compacta, esto los ha vuelto la técnica de extracción de características más usada en reconocimiento del habla [2]. Para la elaboración de un vector, primeramente, se aplica un filtro de pre-énfasis a la señal y posteriormente se divide la misma en tramas y se le aplica una función de ventaneo. El ventaneo sirve para eliminar los bordes de la señal y darle una acentuación a la parte central de la trama para su análisis.[2]

Al obtener la Transformada Discreta de Fourier (DFT) de cada trama se utiliza la amplitud del espectro, y esta información es pasada al dominio de Mel mediante el Banco de Filtros. La escala Mel se basa en mapear entre la frecuencia actual al pitch que percibe, un escucha humano simulado, esta escala es lineal por debajo de 1 kHz y logarítmica por encima de este umbral. Después se obtiene el logaritmo de la señal y finalmente se aplica la Transformada de Coseno Discreta (DCT), de este vector obtenido se toman la cantidad de coeficientes deseados por trama.[2]

V. Redes Neuronales Convolucionales

Dentro de los algoritmos de esta área basados en redes neuronales, en la literatura han sido de particular interés aquellos que primero transforman las señales en imágenes

multicanal. Los algoritmos utilizados en estos y otros trabajos clasifican la ictalidad de las imagenes obtenidas por medio de redes neuronales convolucionales. La operación fundamental en este tipo de red es la convolución. Esta obtiene una imagen de una original mediante combinaciones lineales aprendidas de vecindades de pixeles en la original. La Figura 3.2 muestra la aplicación de varias convoluciones diferentes a una región particular de una imagen.

Una capa típica de una red neuronal convolucional, conocida en la literatura como capa convolucional, consiste en la aplicación secuencial de las siguientes operaciones:

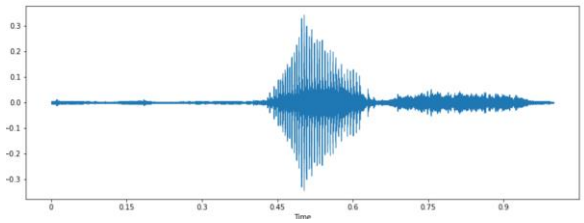
1. Batch-normalization: Esta operación es similar a la estandarización de variables, en la cual, a cada una de las variables independientes, se le resta su promedio y se divide en su desviación estándar en el conjunto observado. La diferencia es que, en el caso de las redes neuronales, realizar este proceso con respecto al conjunto total de datos sería muy costos. Por esta razón se lleva a cabo con respecto al subconjunto de datos que se tiene en cuenta en la iteración actual (conocido como batch). Su aplicación ayuda a transformar la función de perdida en una más circular, acelerando su optimización, y a que cada neurona de la red aprenda más independientemente de las demás [2].
2. Convoluciones: Se aplican, independientemente, varias convoluciones. Se aplica la función de activación ReLU al resultado. [2]
3. Max Pooling: Esta operación reduce la dimensión de la imagen, mediante el emplazo de una región por un ´único pixel con el valor máximo encontrado en la región. Esta reducción permite a la red encontrar características más semánticas en capas más profundas, pues estas ya no observan una imagen sino un resumen pequeño de esta. [2]

VI. Datos de entranmiento y validacion

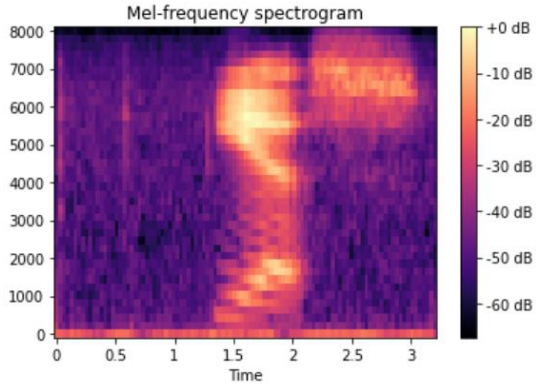
El dataset dado para entrenar la red de convolución cuenta con 65125 archivos de audios, de los cuales 6798 son para validación y el resto para entrenamiento, los cuales de clasifican en 11 clases:

1. Clase ID: 0, clase: yes
2. Clase ID: 1, clase: no
3. Clase ID: 2, clase: up
4. Clase ID: 3, clase: down
5. Clase ID: 4, clase: left
6. Clase ID: 5, clase: right
7. Clase ID: 6, clase: on
8. Clase ID: 7, clase: off
9. Clase ID: 8, clase: stop
10. Clase ID: 9, clase: go
11. Clase ID: 10, clase: silencio/rudio de fondo
12. Clase ID: 11, clase: desconocido

A continuación, se muestra la grafica de audio y del espectrograma para un archivo de audio de cada clase.

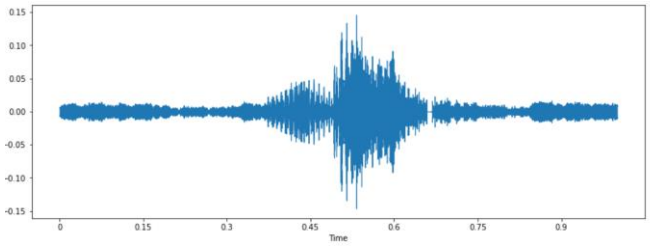


a) Audio

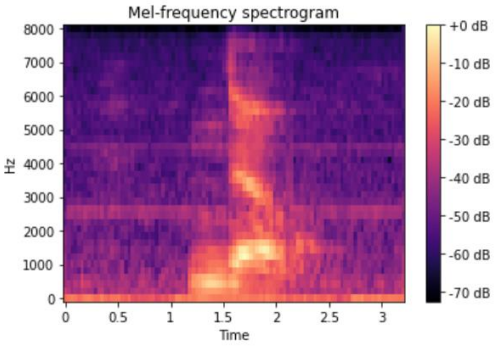


b) Espectrograma

Figura 1. Clase yes

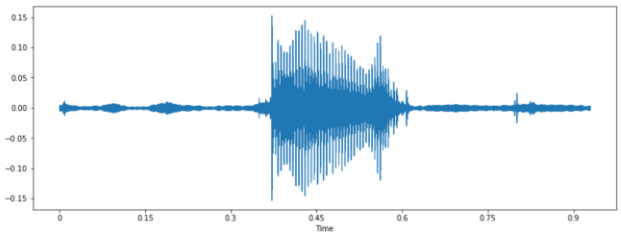


a) Audio

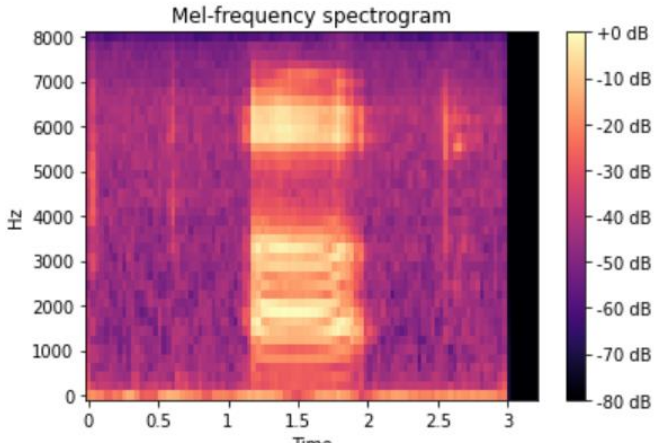


b) Espectrograma

Figura 2. Clase no

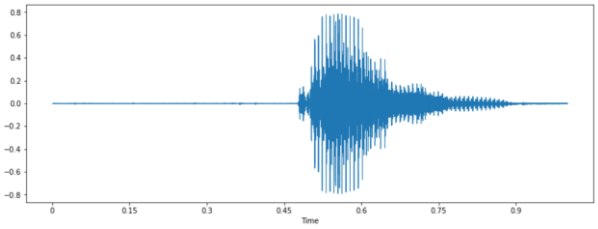


a) Audio

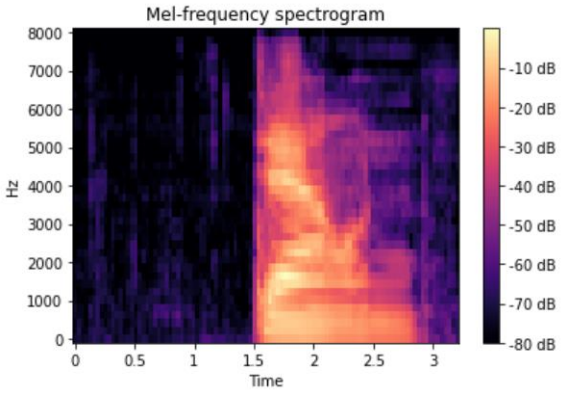


b) Espectrograma

Figura 3. Clase up

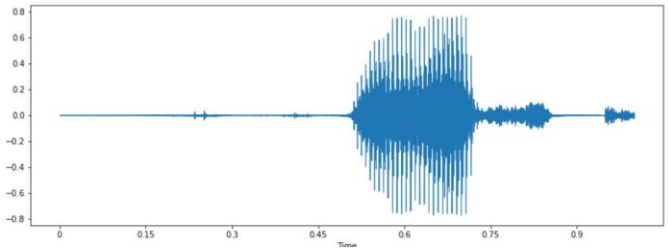


a) Audio

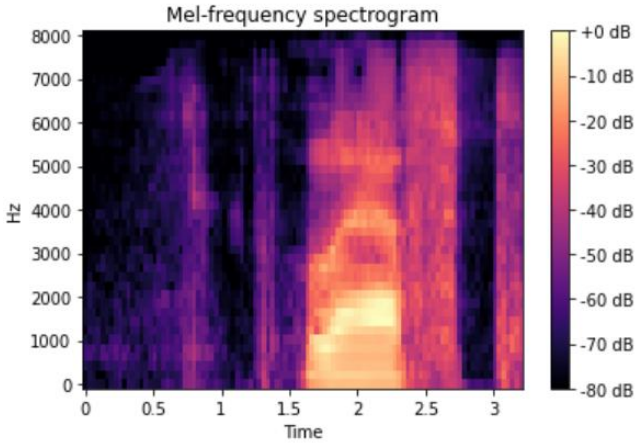


b) Espectrograma

Figura 4. Clase down

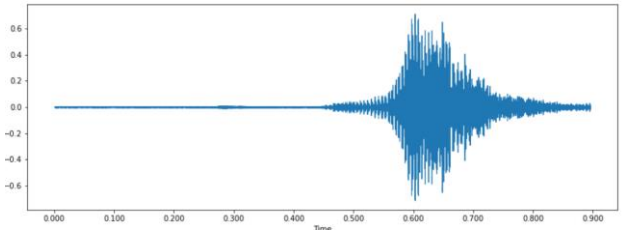


a) Audio

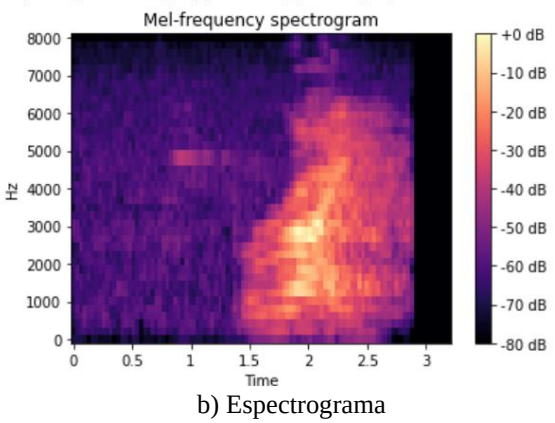


b) Espectrograma

Figura 5. Clase left

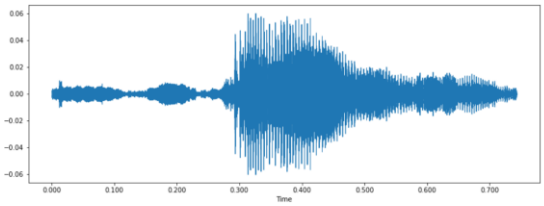


a) Audio

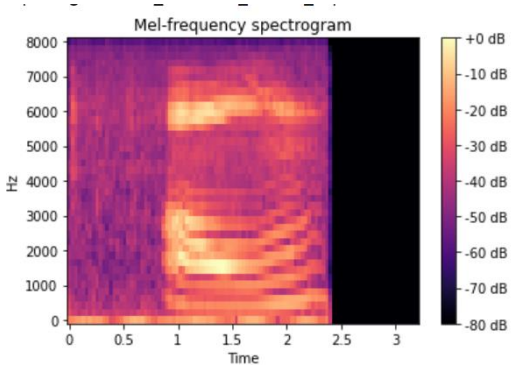


b) Espectrograma

Figura 6. Clase right

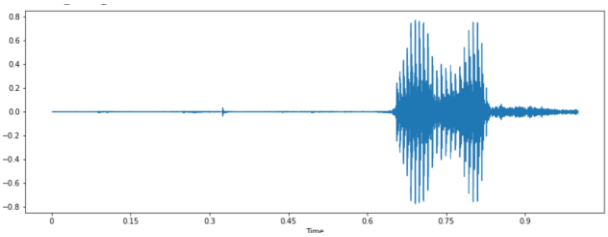


a) Audio

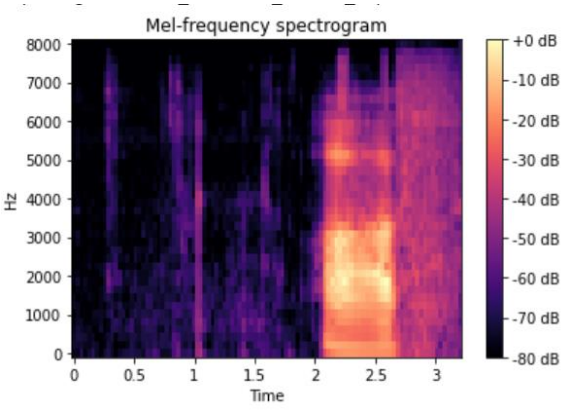


b) Espectrograma

Figura 7. Clase on

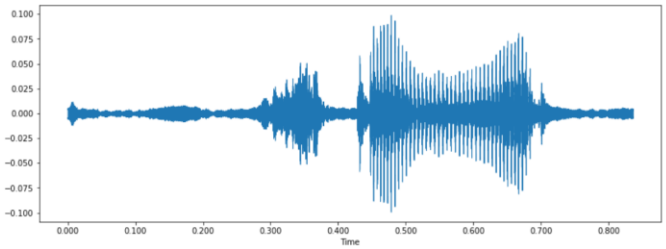


a) Audio

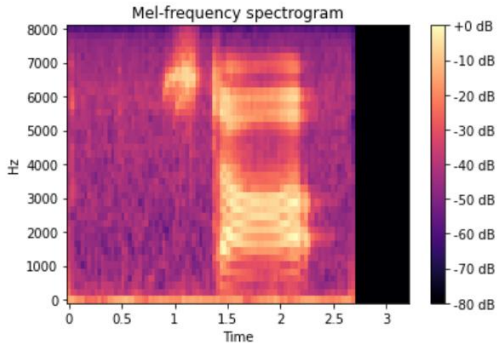


b) Espectrograma

Figura 8. Clase off

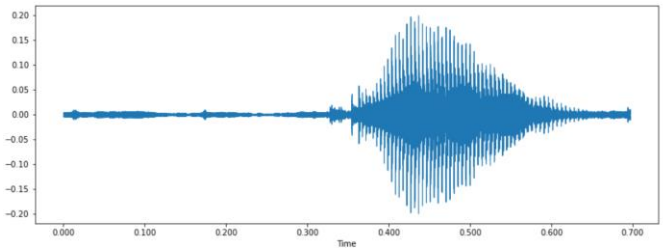


a) Audio

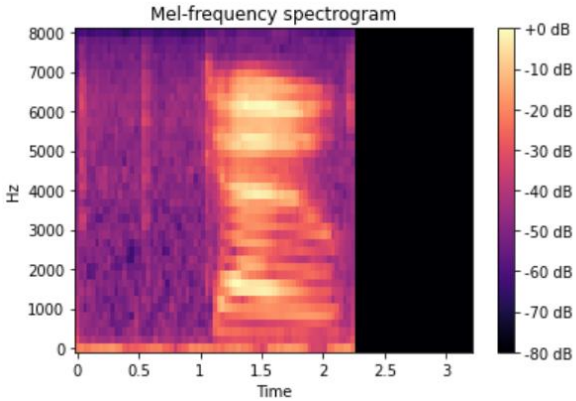


b) Espectrograma

Figura 9. Clase Stop

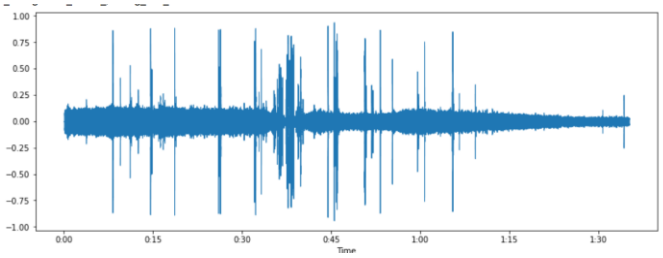


a) Audio

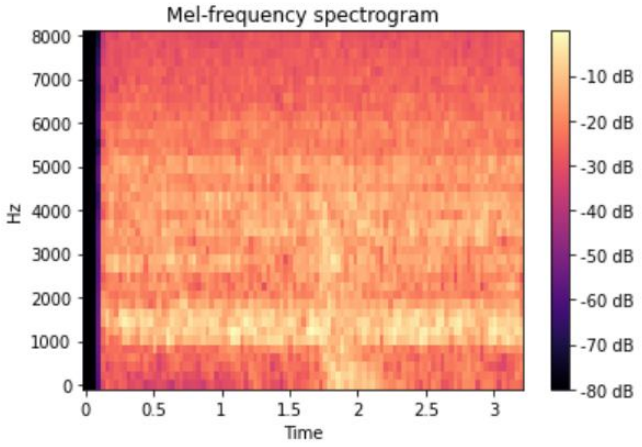


b) Espectrograma

Figura 10. Clase go



a) Audio

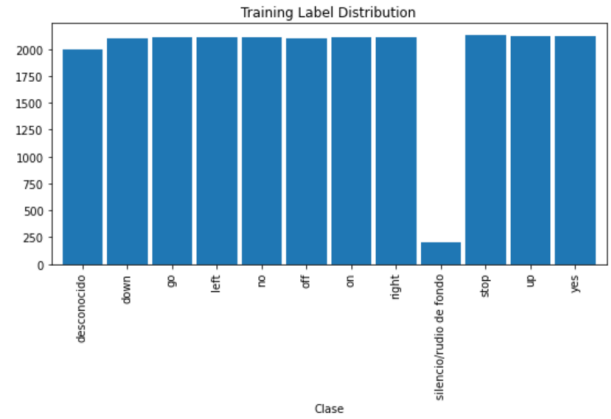


b) Espectrograma

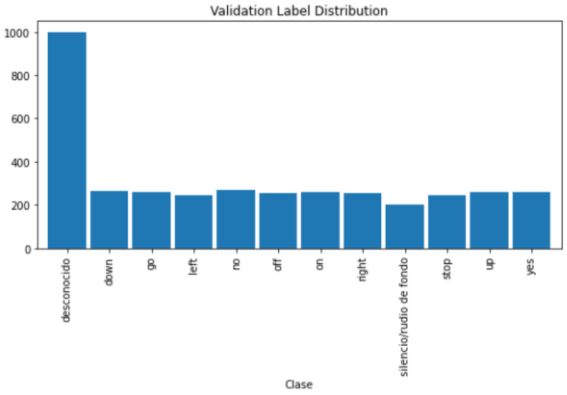
Figura 11. Clase silencio/ruido de fondo

VII. TRATAMIENTO DE LOS DATOS

Debido hay que muchas de datos de la clase desconocido se eliminaron varias muestras al azar para quedar con solo 3000 muestras. Adicionalmente los archivos de audio correspondientes a ruido de fondo se dividieron en muestras de 1s y se distribuyeron homogéneamente entre el conjunto de entrenamiento y de validación. De forma que al final los conjunto de datos quedaron coo se muestra en la figura 12. En la figura 13 se muestra el histograma de del total de los datos entre ambos conjuntos.



a) Conjunto de datos de entrenamiento



b) Conjunto de datos de validación

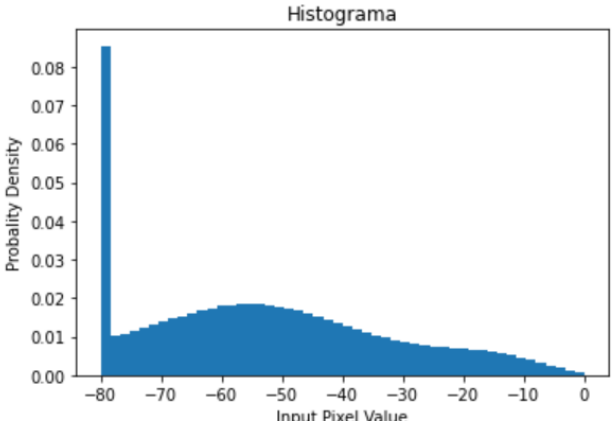


Figura 13. Histograma de conjunto de datos

VIII. ENTRANEMINTO DE RED NEURONAL

En el entrenamiento de la red neuronal las métricas fueron mejorando como se ve en la figura 14 y la figura 15.

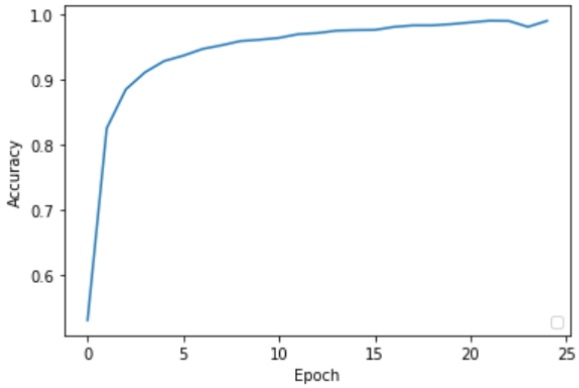


Figura 14. Grafica de accuaracy en el entrenamiento del modelo

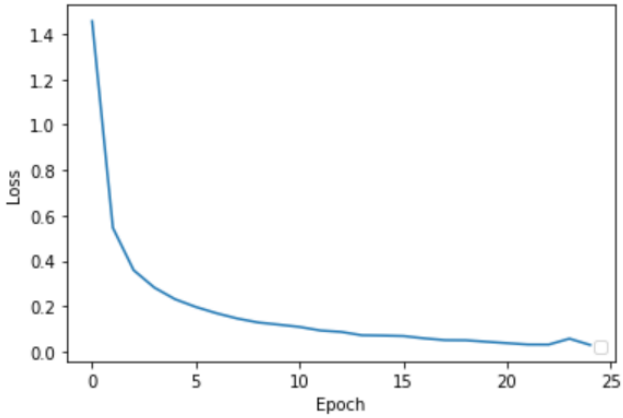


Figura 15. Grafica de Loss en el entrenamiento del modelo.

IX. Evaluacion de red

Para evaluar la red se uso tanto el dataset de entrenamiento y el de validación. En la figura 16 se ve las métricas para el conjunto de entrenamiento y en la figura 17 la matriz de confusión correspondiente a este conjunto. En la figura 18 se ve las métricas para el conjunto de entrenamiento y en la figura 19 la matriz de confusión correspondiente a este conjunto.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2116
1	1.00	0.97	0.98	2105
2	0.99	1.00	0.99	2115
3	1.00	0.98	0.99	2095
4	0.96	1.00	0.98	2106
5	1.00	0.97	0.99	2111
6	1.00	0.95	0.97	2110
7	0.95	1.00	0.97	2101
8	1.00	1.00	1.00	2134
9	0.97	1.00	0.98	2112
10	0.99	1.00	0.99	205
11	0.99	1.00	1.00	2000
accuracy			0.99	23310
macro avg	0.99	0.99	0.99	23310
weighted avg	0.99	0.99	0.99	23310

Figura 16. Métricas del modelo predicción usando los datos de entrenamiento

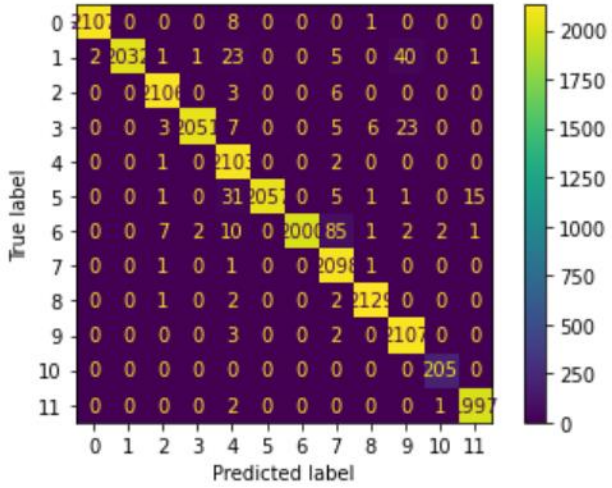


Figura 17. Matriz de confusión del modelo de predicción usando los datos de entrenamiento.

	precision	recall	f1-score	support
0	0.86	0.97	0.91	261
1	0.96	0.86	0.91	270
2	0.66	0.91	0.77	260
3	0.76	0.87	0.81	264
4	0.63	0.98	0.77	247
5	0.84	0.90	0.87	256
6	0.89	0.87	0.88	257
7	0.64	0.97	0.77	256
8	0.66	0.94	0.78	246
9	0.66	0.92	0.77	260
10	0.85	0.06	0.10	199
11	0.96	0.54	0.69	1000
accuracy			0.77	3776
macro avg	0.78	0.82	0.75	3776
weighted avg	0.82	0.77	0.75	3776

Figura 18. Métricas del modelo predicción usando los datos de validación.

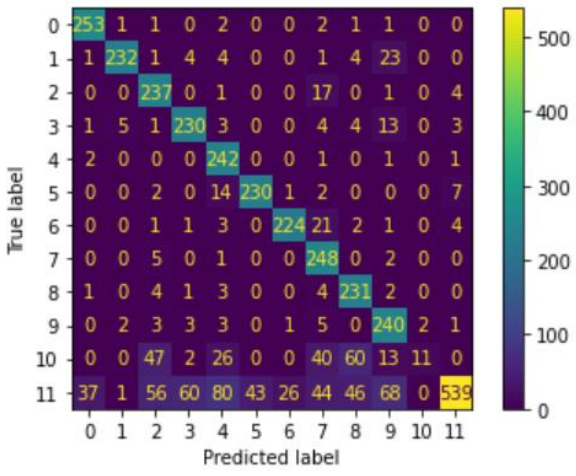


Figura 19. Matriz de confusión del modelo de predicción usando los datos de validación.

Como es de esperarse los resultados son sumamente buenos para el conjunto de entreanamiento aunque no tanto para el de validación. Sin embargo, los valores de precisión, accuary recall y f1 rondan el 0.7 lo cual es un valor bastante aceptable. El valor de accuaracy indica que el modelo acierta la mayor parte de las veces. El de precisión indica que es mayor la cantidad de verdaderos positivos que de falsos positivos. Un buen recall quiere deir que hay pocos falsos negativos. En general se puede afirmar que la predicción del modelo es aceptable, sin embargo, habría que considerar la aplicación para reafirmar esto. Por otro lago, hay que ocnsidera que el costo computacional del modelo es najo, y es capaz de procesar un espectrograma cada 30ms.

X. CONCLUSIONES

El uso de redes neuronales es adecuado para el reconocimiento de audios. Se pueden manipular los parámetros de la misma para buscar hacerla mas precisa en caso de que se requiera.

REFERENCIAS

[1] Cerón Uribe, J. (2019). Redes convolucionales para la predicción de convulsiones epilépticas. La transformada de Fourier de tiempo corto y el impacto de sus hiperparámetros en la calidad de las predicciones. Universidad de los Andes.

[2] G.Martinez y G.Aguilar “Reconocimiento de voz basado en MFCC, SBC y Espectrogramas” en Ingeneous Revista de Cinecia y Tecnologia. Universidad Politecnica Saleciana de Ecuador.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published