La Mecatrónica en México, Mayo 2022, Vol. 11, No. 2, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

Clasificación de Señales de Audio FSDD Utilizando Redes Neuronales Convolutivas

Hernández Montejano Carlos Oliver, González Huerta Rodrigo,
Tovar Arriaga Saúl (✉)

Facultad de Ingeniería, Universidad Autónoma de Querétaro, México

✉saul.tovar@uaq.mx (corresponding author)

Resumen

Free Spoken Digit Dataset (FSDD) de Zohar Jackson y col. [1] es una base de datos inspirado en el famoso dataset MNIST de Yann LeCun [2]. Está conformada por grabaciones de audio de los digitos cero a nueve con diferentes personas, llevando la exploración al tratamiento de señales de audio. En este artículo se desarrollaron dos modelos de redes neuronales convolutivas con la finalidad de clasificar las señales de audio en su número correspondiente entre cero y nueve. El objetivo es comparar los resultados de ambos modelos, para identificar la opción que ofrezca un mejor desempeño en la tarea de clasificación. Uno de los modelos se centra en los patrónes en los datos de las señales, el segundo en las características que aparecen en imágenes de espectrograma de cada señal y posteriormente ambos realizan la clasificación. Se obtuvo 87.6 % de exactitud con la clasificación de la señal de audio y 94.7 % con la clasificación de imágenes de espectrograma. Además se validan ambos modelos con un conjunto de grabaciones propias tomadas en diferentes condiciones de grabación.

Palabras clave: Clasificación, Señales de Audio, FSDD, Espectrogramas, Redes Neuronales Convolutivas.

Abstract

Free Spoken Digit Dataset (FSDD) from Zohar Jackson et al. [1] is an inspiring database on the famous MNIST dataset by Yann LeCun [2]. It is made up of audio recordings of the digits zero to nine with different people, performing the exploration to the treatment of audio signals. In this article, two models of convolutional neural networks were developed in order to classify the audio signals in their corresponding number between zero and nine. The objective is to compare the results of both models, to identify the option that offers a better performance in the classification task. One of the models focuses on the patterns in the signal data, the second on the features that appear in spectrogram images of each signal, and then both perform the classification. An 85.2% accuracy was obtained with the audio signal classification and 97% with the spectrogram image classification. In addition, both models are validated with a set of own recordings taken under different recording conditions.

Keywords: Classification, Audio Signals, FSDD, Spectrograms, Convolutional Neural Networks.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

1. Introducción

La clasificación de audio es una tarea importante y de alto impacto en la actualidad. Sistemas de reconocimiento de voz, asistentes inteligentes, reconocimiento de emociones, género o sonidos hambientales son algunas aplicaciones en las que los sistemas de clasificación de sonidos son de utilidad tanto para sistemas médicos de inteligencia artificial o sistemas de conversación.

En cuanto a sistemas de clasificación de audio, el aprendizaje profundo ha demostrado ser mas competente que las técnicas de aprendizaje automático, se han utilizado para diversas tareas que involucran la clasificación de audio [3][4], algunas de ellas como traducción automática [5], subtituló de imágenes [6] y detección de eventos de sonido [7][8].

En el campo del aprendizaje profundo, existen diferentes tipos de redes neuronales que se desempeñan mejor que otras en diferentes tareas. Tal es el caso de las redes neuronales convolutivas, las cuales tienen un gran desempeño en tareas de clasificación de imágenes y audio [9][15] .

A diferencia de los humanos, que procesan el sonido de forma natural en su forma continua, las computadoras implementan modelos matemáticos para hacer representaciones discretas del sonido. Esta discretización del sonido es denominada “Digitalización”. La digitalización traduce las señales del mundo físico en secuencias de números que las computadoras pueden aceptar como entrada y posteriormente procesar de alguna manera conveniente. La conversión de una señal analógica en digital requiere muestrear la primera en instantes de tiempo específicos [10-12].

Mediante la aplicación de una transformada matemática, llamada transformada de Fourier, cualquier señal de audio puede descomponerse en un conjunto de ondas periódicas. Esta es una herramienta matemática valiosa especialmente para analizar señales no periódicas como los sonidos que producimos los seres humanos, por ejemplo, la voz [12].

La transformada de Fourier también se puede calcular en breves ventanas de tiempo superpuestas; esto se llama transformada de Fourier de corta duración. Luego, para cada uno de estos pedazos, la magnitud del espectro de frecuencia de la señal se puede obtener y trazar como un gráfico 3D, con el tiempo, la frecuencia y la magnitud como componentes, pero con la sutileza de que la magnitud queda representada por colores en lugar de utilizar otra dimensión espacial. A este tipo de representación se le llama espectrograma [12].

En términos más simples, los espectrogramas son representaciones visuales del sonido. Un espectrograma muestra qué frecuencias componen una señal de sonido y también la forma en como varían con el tiempo. De los espectrogramas, se pueden extraer características de una señal de audio. Las características son solo información significativa que se utiliza para diferenciar los distintos tipos de señales. Al mismo tiempo, los espectrogramas ayudan a descartar información sin sentido que en una grabación de audio podría ser solo ruido de fondo [13][15].

Existen algunos trabajos interesantes que utilizan estas mismas bases para realizar clasificación de señales de audio y reconocimiento de voz, por ejemplo: Giobergia [16] con modelos de bosques aleatorios y maquinas de soporte vectorial y Nasr [17] con una red neuronal profunda, clasifican los dígitos hablados de la base de datos FSDD.

Otros como Sharmin [18] y Das [19] hacen uso de las redes neuronales convolutivas para clasificación de dígitos hablados en idioma Bengalí, ya que en Bangladesh ha sido difícil el despliegue y uso de aplicaciones que funcionan a través de la voz.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

A lo largo de este artículo, se propone una comparativa entre dos modelos de redes neuronales convolutivas diferentes. Uno dedicado al análisis e identificación de características de los datos de la señal de audio, mientras que el otro se enfoca en la obtención de características de imágenes de espectrograma obtenidos a través de la aplicación del espectrograma de Mel. Con esta implementación el objetivo es averiguar cuál de las dos técnicas podría ser más efectiva en términos de exactitud, tiempo y complejidad. Además se verifica la eficacia de los modelos poniéndolos a prueba con grabaciones de audio propias y así poder observar su comportamiento en condiciones de grabación diferentes a las utilizadas con la base de datos de entrenamiento.

2. Metodología

La metodología utilizada en el desarrollo de este proyecto es representada por la Figura 1. Primeramente, se estudió el comportamiento de las señales de audio, se tomaron varias muestras de las señales y se plasmaron en un gráfico para apreciar la actividad del sonido en una línea de tiempo. También se verificó la duración de cada una de las señales a través de una gráfica, así como el número de muestras que se tienen por cada una de las clases en el argumento de decisión.

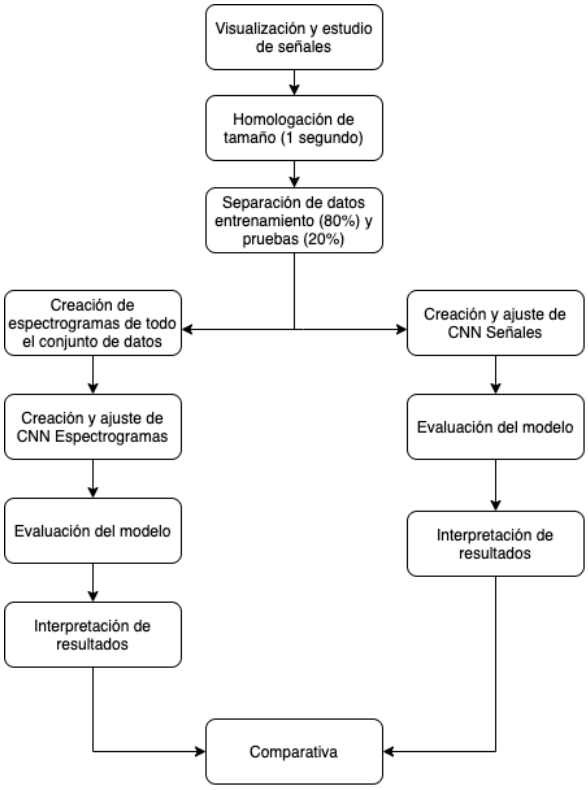

Figura 1. Diagrama de metodología.

Una vez estudiados estos datos se procedió a homologar el tiempo de cada audio, se estableció una duración de 1 segundo para todas las muestras sin perder información en la señal de audio. Posteriormente se realizó la separación en dos conjuntos de datos, 80% para entrenamiento de los modelos y 20% para pruebas. A partir de este punto, la metodología se separa para cubrir el modelo que trabajará con los datos de la señal y el que trabajará con las imágenes de espectrograma.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

Por un lado, se procede a la creación y ajuste de un modelo de red neuronal convolucional que recibirá como parámetros de entrada las señales de audio, posteriormente se evaluará el modelo con el conjunto de pruebas y por último se realiza la interpretación de los resultados, dónde se observará el desempeño y la capacidad del modelo para clasificar correctamente las señales de audio.

Por otro lado, se comienza por la transformación de las señales de audio en imágenes de espectrograma, para realizar esto, se aplica una técnica conocida como Mel Frecuency Cepstral Coefficients (MFCC) o espectrograma de Mel. El MFCC es una técnica para extracción de características de señales de audio dónde se aplica la Transformada Discreta de Fourier (DFT) a ventanas de la señal, tomar la magnitud para después deformar las frecuencias en una escala de Mel y por último se aplica la Transformada de Coseno Discreta (DCT) inversa [20].

De forma más detallada, el proceso MFCC para extracción de características consta de los siguientes pasos: Pre-énfasis, que se refiere al filtrado que enfatiza las frecuencias más altas. Su propósito es equilibrar el espectro de sonidos sonoros que tienen una caída pronunciada en la región de alta frecuencia. Este filtro está dado por la Ecuación 1, dónde el valor de *b* controla la pendiente del filtro y suele estar entre 0.4 y 1.0 [20].

$$H(z) = 1 - bz^{-1} \tag{1}$$

El bloqueo de tramas y ventanas, se utiliza para obtener características acústicas y estables examinando la señales de audio en periodos de tiempo cortos. Las ventanas suelen ser de 20ms con un avance de 10 *ms*, esto permite obtener una buena resolución espectral de los sonidos y resolver características temporales significativas. Todo este procedimiento se hace para mejorar los armónicos, suavizar los bordes y reducir el efecto de borde mientras se toma la DFT en la señal [20].

Cada marco de ventana se convierte a espectro de magnitud aplicando la Transformada Discreta de Fourier, este espectro se obtiene con la Ecuación 2, dónde *N* es el número de puntos usados para calcular la DFT [20].

$$\boldsymbol{X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}; 0 \leq k \leq N-1} \tag{2}$$

El espectro de Mel se calcula pasando la señal DFT a través de un conjunto de filtros de paso de banda. Un Mel es una unidad de medida que se basa en la frecuencia percibida por los oídos humanos. La escala de Mel es un espacio aproximado de frecuencia lineal que se encuentra por debajo de 1 kHz y un espaciado logarítmico por encima de 1 kHz y puede ser expresada como en la Ecuación 3 dónde *f* denota la frecuencia física en *Hz* y f_{Mel} denota la frecuencia percibida [20].

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

El filtro utilizado mas comúnmente es el triangular, en algunos otros casos puede encontrarse el Hanning o Hamming [21]. Una representación de estos filtros triangulares se puede apreciar en la Figura 2 [20].

El espectro de Mel se calcula obteniéndolo del espectro de magnitud ($X(k)$), multiplicando el espectro de magnitud por cada uno de los filtros triangulares de ponderación de Mel, esto se aprecia

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

en la Ecuación 4 dónde *M* es el número de filtros triangulares de ponderación de Mel, $H_m(k)$ es el peso dado al k^{th} contenedor de espectro de energía que contribuye a la m^{th} banda de salida [20], la cual se expresa como en la Ecuación 5 dónde m va de 0 a M-1 [20].

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 H_m(k)] \, ; 0 \leq m \leq M-1 \tag{4}$$

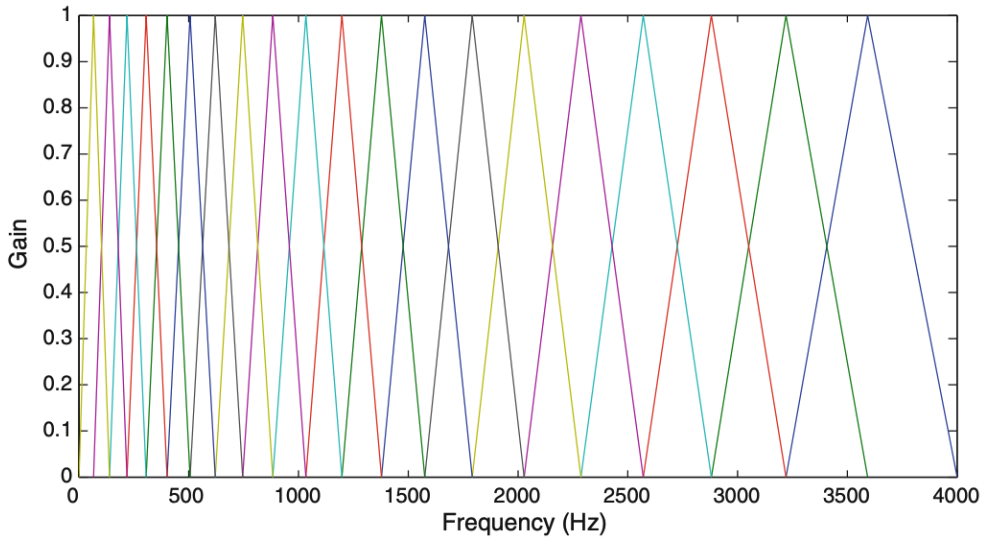

Figura 2. Banco de filtros de Mel tomado de [20].

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{2(k - f(m-1))}{f(m) - f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \dfrac{2(f(m+1) - k)}{f(m+1) - f(m)}, & f(m) < k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{5}$$

La Transformada de Coseno Discreta (DCT) se aplica a los coeficientes de frecuencia de Mel transformados, esto produce un conjunto de coeficientes cepstrales. Antes de calcular DCT, el espectro de Mel generalmente se representa en una escala logarítmica. Esto da como resultado una señal en el dominio cepstral con un pico de quefrecuencia correspondiente al tono de la señal y una serie de formantes que representan picos de baja quefrecuencia [20]. La mayoría de la información de la señal se representa en los primeros coeficientes MFCC, por lo que se puede extraer solamente estos coeficientes ignorando o truncando los componentes del DTC de orden superíor. El cálculo del MFCC se realiza con la Ecuación 6 dónde $c(n)$ son los coeficientes cepstrales y C es el número de MFCC. Tradicionalmente se utilizan de 8 a 13 coeficientes cepstrales. El coeficiente cero a menudo se excluye ya que representa la energía logarítmica promedio de la señal de entrada y esta transporta muy poca información específica [20].

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(s(m)) \cos\left(\frac{\pi n(m-0.5)}{M}\right); n = 0,1,2, \dots, C-1 \quad (6)$$

Los coeficientes cepstrales generalmente se denominan funciones estáticas, ya que solo contienen información de un marco determinado. La información adicional de la señal se obtiene calculando la primera y segunda derivada de los coeficientes cepstrales [22][23][24]. La primera derivada se le conoce como coeficientes delta, el cual informa sobre la velocidad del habla, y a la segunda como coeficientes delta-delta, estos brindan información similar a la aceleración del habla. La definición más común para calcular el parámetro dinámico es con la Ecuación 7 dónde $c_m(n)$ denota la función m^{th} para el marco de tiempo n^{th}, ki es el peso i^{th} y T es el número de tramas utilizadas para el cálculo. Generalmente T se toma como. Los coeficientes delta-delta se calculan tomando la derivada de primer orden de los coeficientes delta [20].

$$\Delta \boldsymbol{c_m}(\boldsymbol{n}) = \frac{\sum_{\boldsymbol{i}=-\boldsymbol{T}}^{\boldsymbol{T}} \boldsymbol{k_i c_m}(\boldsymbol{n}+\boldsymbol{i})}{\sum_{\boldsymbol{i}=-\boldsymbol{T}}^{\boldsymbol{T}} |\boldsymbol{i}|} \quad (7)$$

Una vez se realiza la transformación de señales a imágenes de espectrograma se procede con la creación y ajuste de la red neuronal convolucional que trabajará con las imágenes para realizar la clasificación. Posteriormente se evalúa el modelo con el conjunto de pruebas y se interpretan los resultados para verificar su desempeño y capacidad de clasificación.

Por ultimo se realiza una comparativa entre ambos modelos para identificar cuál se desempeña mejor en la tarea de clasificación, considerando tiempo y recursos. Además se pretende evaluar la capacidad de clasificación del modelo utilizando un conjunto de pruebas con grabaciones propias, esto permitirá observar que tan eficiente es el modelo aun con datos en condiciones grabación distintas y que además no fueron utilizadas en el entrenamiento.

3. Datos

La base de datos Free Spoken Digit Dataset (FSDD) de Zohar Jackson y col. [1] está formada por 3,000 grabaciones de dígitos hablados de 0 a 9, provenientes de 6 sujetos. En cierto sentido, es análoga al dataset MNIST [2], utilizado para clasificar dígitos escritos. Otros datos específicos de esta base de datos se muestran a continuación:

1. Formato de los archivos WAV a 8 kHz.
2. Las grabaciones están recortadas de forma que tienen un silencio mínimo al principio y al final.
3. Duración variable. 2.28 segundos es la duración del archivo más largo y 0.14 la del archivo más corto.
4. Pronunciación en inglés.
5. Los archivos se nombran de la siguiente manera: {digito}_{sujeto}_{índice}.

En la Figura 3 se pueden visualizar nueve muestras de las grabaciones de la base de datos FSDD tomadas de forma aleatoria y en la Figura 4 se puede visualizar su equivalente en espectrogramas de Mel.

A manera de prueba y con la finalidad de demostrar la capacidad de clasificación de los modelos, se utilizó una base de datos con grabaciones propias, con base en las mismas características de la base de datos FSDD. Esto se realiza con la finalidad de comprobar la eficiencia de la red al intentar clasificar señales de audio en condiciones de grabación distinta a las utilizadas en el entrenamiento. En la Figura 3 se observan las señales de audio de las grabaciones propias y en la Figura 4 su equivalente con espectrogramas de Mel.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

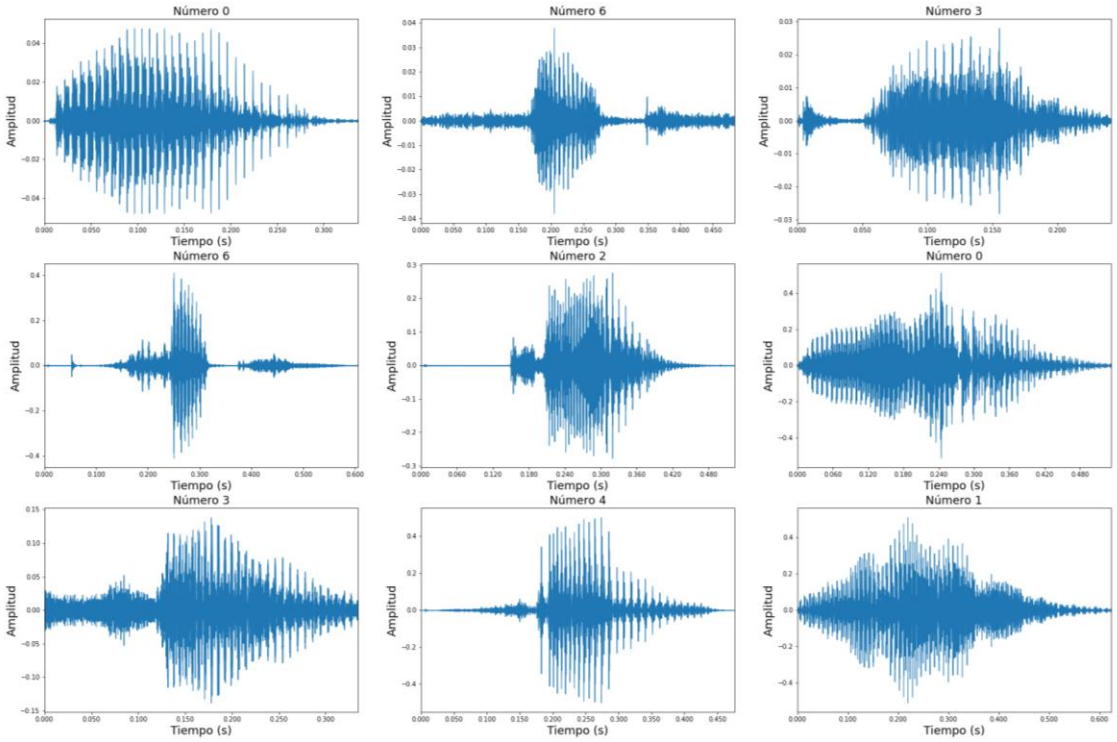

Figura 3. Muestras de grabaciones de audio de la base de datos FSDD.

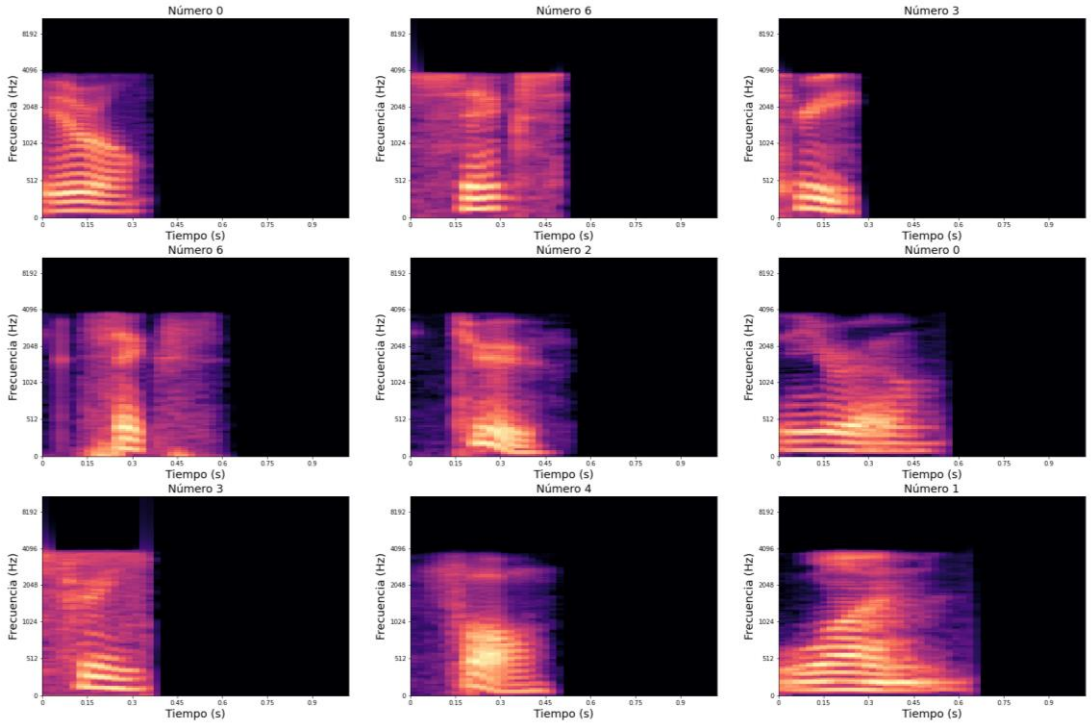

Figura 4. Espectrogramas de Mel obtenidas de las grabaciones de audio FSDD.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

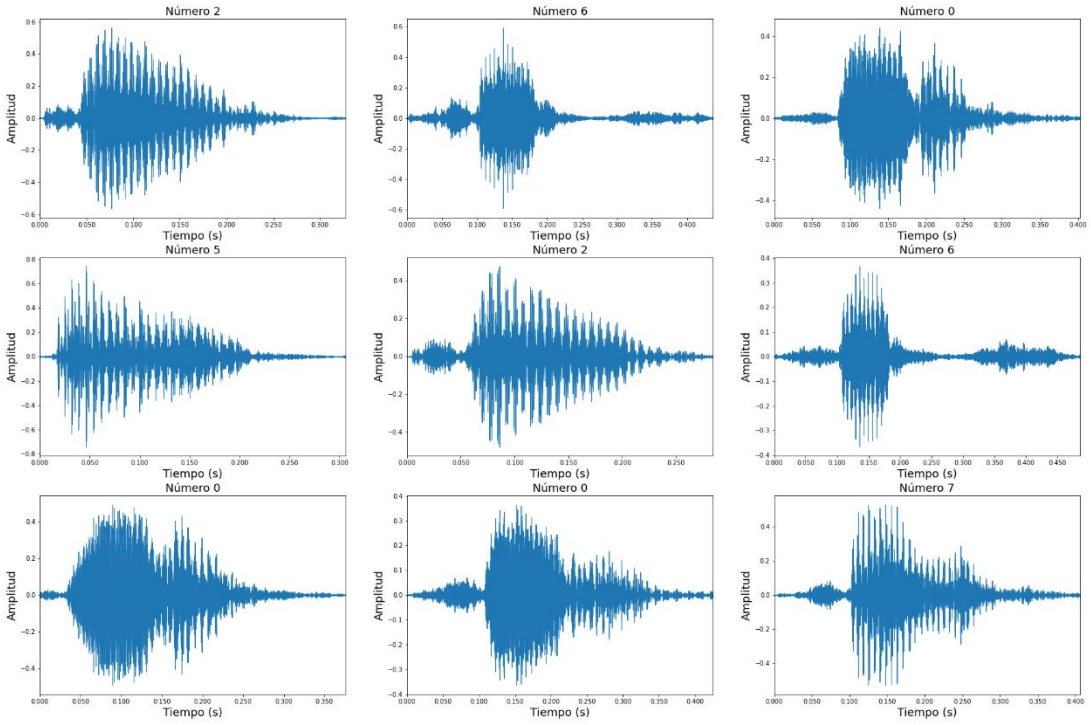

Figura 5. Muestras de grabaciones de audio propias.

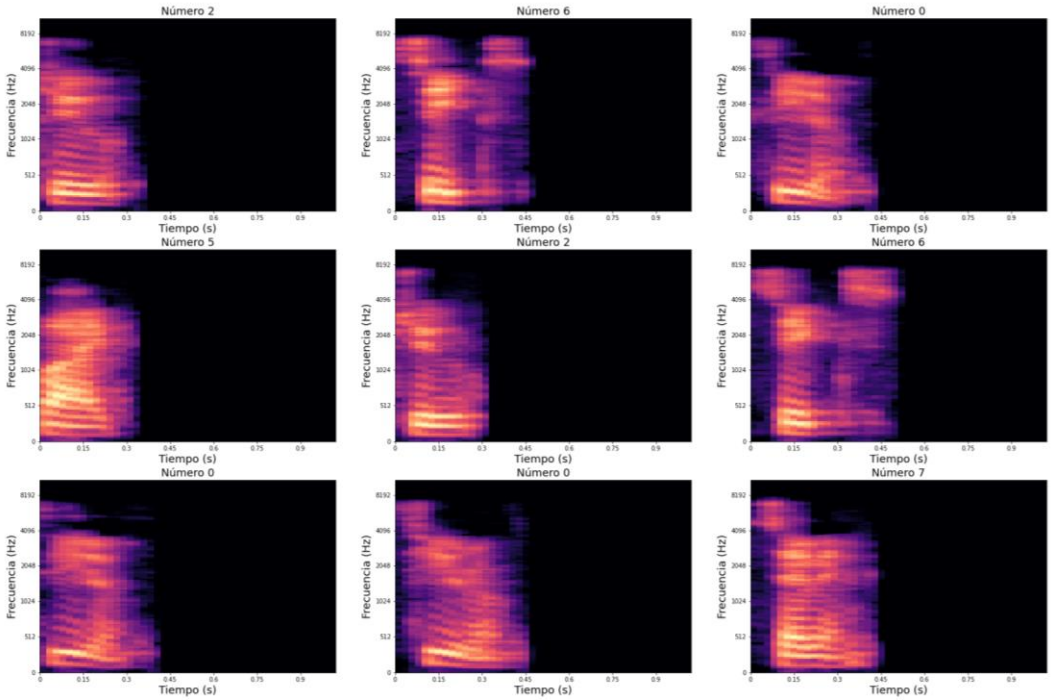

Figura 6. Espectrogramas de Mel obtenidas de las grabaciones propias.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

4. Modelos

4.1 CNN Señales.

Se trabajó con dos modelos diferentes de redes neuronales convolucionales, uno específico para trabajar directamente con los datos de las señales de audio y otro con las imágenes de espectrograma. Para la red convolucional que clasifica los datos de la señales de audio se trabajó una topología sencilla la cual se puede apreciar en la Figura 7 que consta de la capa de entrada, la cual recibirá como entrada la señal de audio recortada a 1 minuto, seguida de dos capas de consolación de una dimensión (Conv1D) y un Dropout de 0.5, una capa de BatchNormalization, un MaxPooling1D, una capa Flatten, un Dropout de 0.3, una capa Densa con activación ReLU y la capa de salida con activación Softmax.

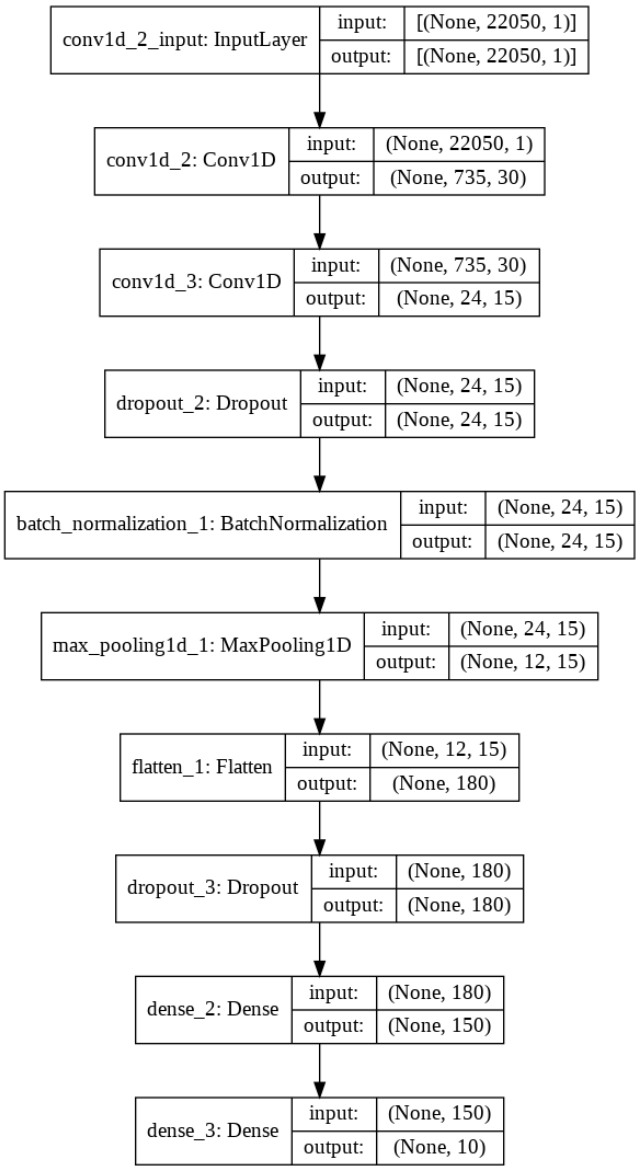

Figura 7. Topologia de CNN Señales.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

4.2 CNN Espectrogramas.

Para la parte de clasificación con espectrogramas se realizó una red neuronal convolucional un poco diferente a la que trabaja con la señales. Para este modelo se utilizó una topología como en la Figura 8, con una capa de entrada que recibe la imagen de espectrograma, seguida de una capa de convolución de dos dimensiones (Conv2D), un MaxPooling2D y un Dropout de 0.25. Se añade una combinación más de Conv2D, MaxPooling2D y Dropout de 0.25, a esto le sigue una capa Flatten, una capa Densa con activación ReLU y la capa de salida con activación Softmax.

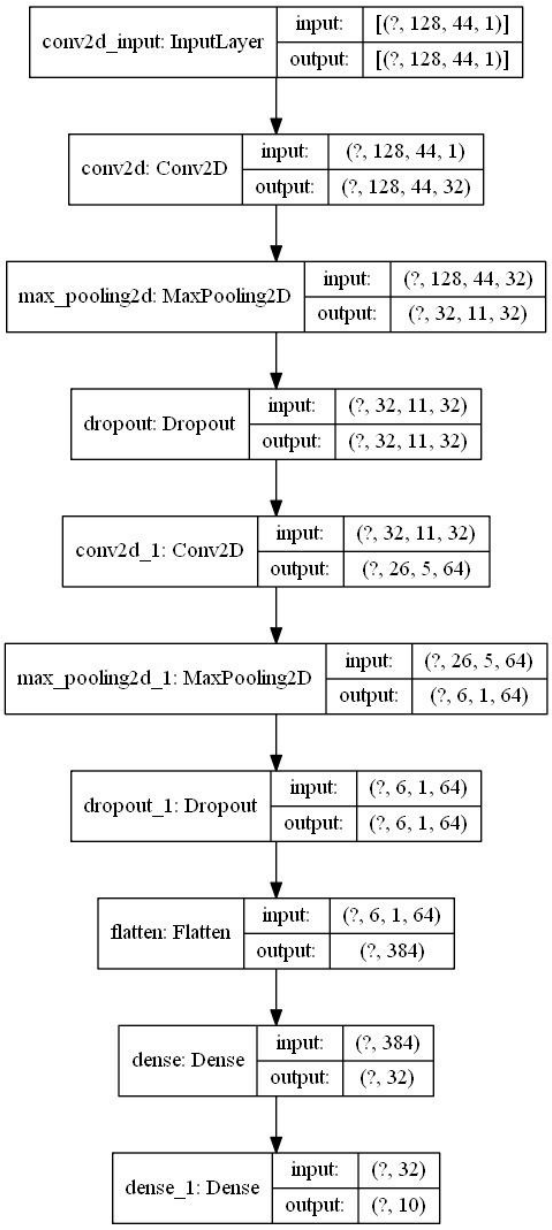

Figura 8. Topologia de CNN Espectrogramas.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

5. Resultados

En los resultados de entrenamiento con la CNN señales, utilizando los datos de la Tabla 1, se puede apreciar que tanto la exactitud y la perdida de la validación va muy cerca del entrenamiento, lo cual indica que la red generaliza bien pero muestra un pequeño *Underfitting*.

Tabla 1. Conjuntos de datos.

Conjunto	Número de muestras	Porcentaje
Entrenamiento	1,500	50%
Validación	500	17%
Pruebas	1,000	33%
Total	3,000	100%

La red se entrenó durante 500 epocas y con un tiempo de entrenamiento aproximado de 30 minutos alcanzó una exactitud del 87.6%, las demás métricas de evaluación se pueden apreciar en la Tabla 2. Las gráficas de entrenamiento se pueden apreciar en la Figura 9. En la Figura 10 se pueden visualizar los resultados de la evaluación, en la matriz de confusión podemos apreciar que la mayoría de los dígitos son clasificados correctamente, algunos que parecen mantener similitud en la pronunciación son los que pueden mostrar algunos errores al momento de ser clasificados. El reporte de clasificación de todas las clases se puede apreciar en la Tabla 3. De forma general el desempeño de la red parece ser considerablemente bueno.

Tabla 2. Resultados FSDD conjunto de pruebas CNN Señales.

Métrica	Porcentaje
Exactitud	87.6%
Precisión	0.88
Recall	0.88
F1	0.88

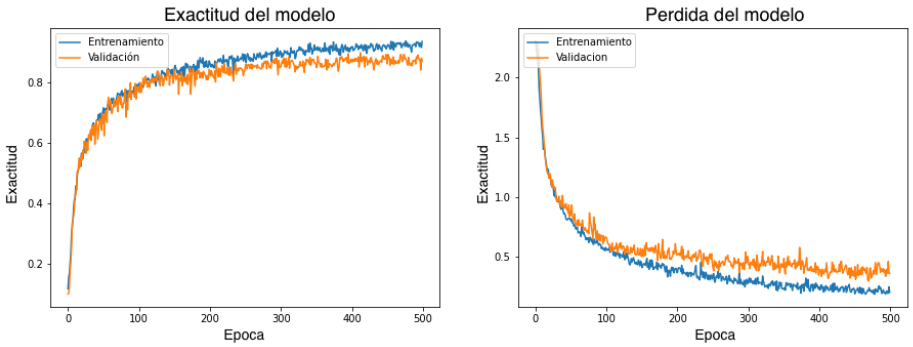

Figura 9. Graficas de entrenamiento y validación CNN Señales.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

Figura 10. Matriz de confusión conjunto de pruebas FSDD CNN Señales.

Con la CNN de espectrogramas se obtuvieron resultados aún mejores, alcanzando una exactitud del 94.7% entre las demás métricas que se muestran en la Tabla 4, sin embargo la pérdida de la validación se dispara durante casi todo el entrenamiento (Figura 11). A partir de esto se puede observar que la red comienza a tener un sobre entrenamiento, siendo capaz de identificar las diferentes muestras de la base de datos FSDD con una exactitud alta. De la gráfica de exactitud en la Figura 11 podemos deducir que la red generaliza de manera excelente.

La evaluación del modelo se puede observar en la Figura 12 con la matriz de confusión, los aciertos son mayores que en el modelo anterior, en la Tabla 4 y en la Tabla 5 también se puede apreciar que los valores en las métricas son mucho mas altos.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

Tabla 3. Reporte de clasificación FSDD conjunto de pruebas CNN Señales.

	Precisión	Recall	F1-Score	Soporte
0	0.94	0.97	0.96	101
1	0.89	0.92	0.91	101
2	0.97	0.80	0.88	96
3	0.73	0.82	0.77	98
4	0.85	0.96	0.90	109
5	0.84	0.83	0.83	104
6	0.85	0.92	0.88	102
7	0.86	0.93	0.90	87
8	0.96	0.82	0.88	98
9	0.92	0.79	0.85	104
Exactitud			0.88	1000
Promedio Macro	0.88	0.88	0.88	1000
Peso promedio	0.88	0.88	0.88	1000

Tabla 4. Resultados conjunto de pruebas FSDD CNN Espectrogramas.

Métrica	Porcentaje
Exactitud	94.7%
Precisión	0.95
Recall	0.95
F1	0.95

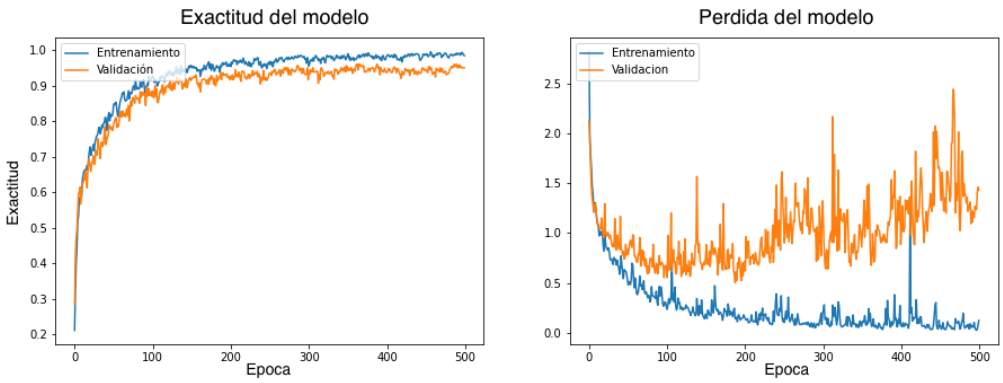

Figura 11. Graficas de entrenamiento y validación CNN Espectrogramas.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

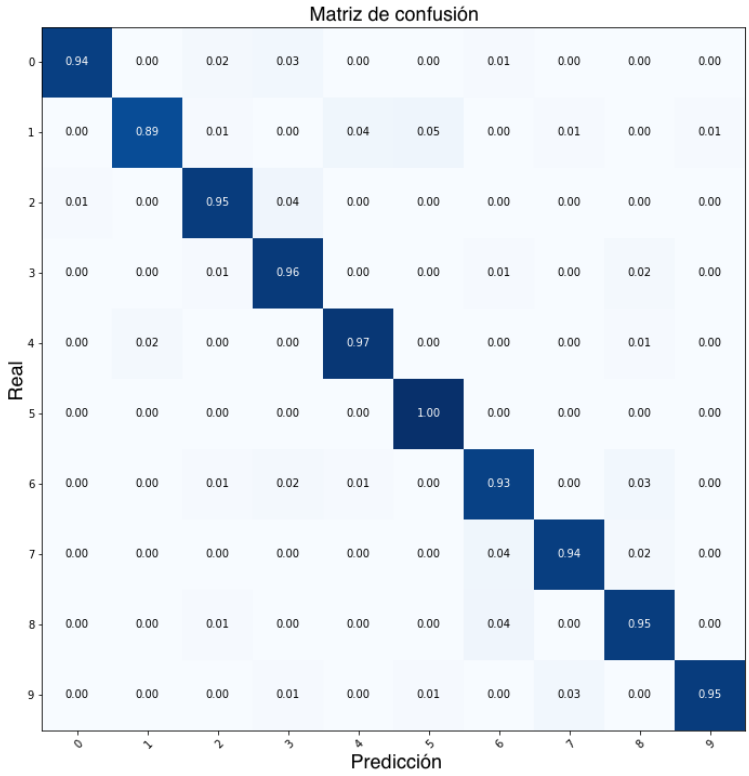

Figura 12. Matriz de confusión conjunto de pruebas FSDD CNN Espectrogramas.

Tabla 5. Reporte de clasificación FSDD conjunto de pruebas CNN Espectrogramas.

	Precisión	Recall	F1-Score	Soporte
0	0.99	0.94	0.96	99
1	0.98	0.89	0.93	109
2	0.94	0.95	0.95	100
3	0.90	0.96	0.93	94
4	0.95	0.97	0.96	105
5	0.94	1.00	0.97	91
6	0.90	0.93	0.92	102
7	0.96	0.94	0.95	104
8	0.92	0.95	0.93	96
9	0.99	0.95	0.97	100
Exactitud			0.95	1000
Promedio Macro	0.95	0.95	0.95	1000
Peso promedio	0.95	0.95	0.95	1000

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

5.1 Grabaciones propias.

Se realizó la evaluación de ambos modelos CNN Señales y CNN Escalogramas con el conjunto de grabaciones propias, en estas pruebas los resultados no fueron relevantes ya que la exactitud de la clasificación se encuentra por debajo del 60% en ambos modelos. Con la CNN señales se alcanzó una exactitud de 55%, mientras que con la CNN espectrogramas solamente se alcanzó el 49%. En la Tabla 6 y la Tabla 7 se encuentran los resultados de evaluación de ambos modelos y en la Figura 13 y Figura 14 se puede visualizar la matriz de confusión de cada uno de ellos. El detalle de las métricas de evaluación para cada una de las clases se encuentra en la Tabla 6 y Tabla 7 respectivamente.

Tabla 6. Resultados conjunto de grabaciones propias CNN Señales.

Métrica	Porcentaje
Exactitud	55.0%
Precisión	0.63
Recall	0.55
F1	0.55

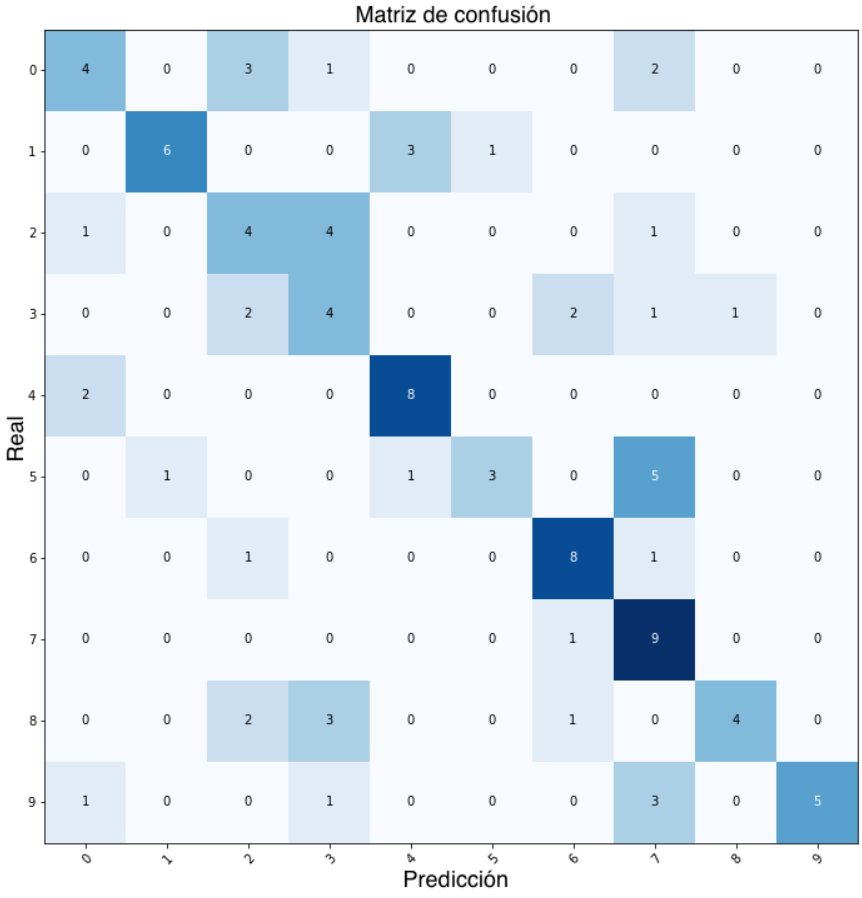

Figura 13. Matriz de confusión conjunto de grabaciones propias CNN Señales.

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

Tabla 7. Resultados conjunto de grabaciones propias CNN Espectrogramas.

Métrica	Porcentaje
Exactitud	49.0%
Precisión	0.55
Recall	0.49
F1	0.48

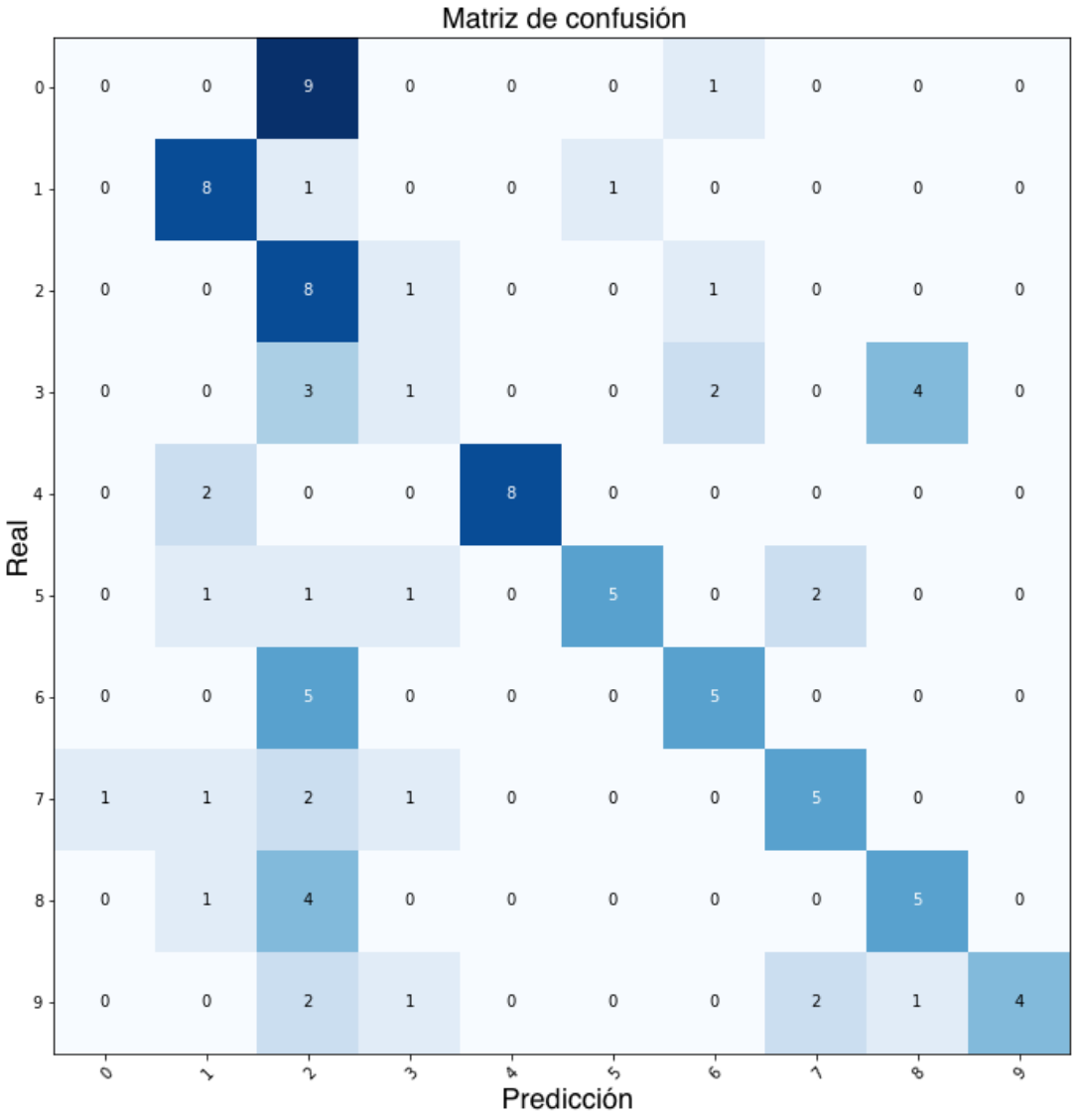

Figura 14. Matriz de confusión conjunto de grabaciones propias CNN Espectrogramas.

Tabla 8. Reporte de clasificación conjunto de grabaciones propias CNN Señales.

	Precisión	Recall	F1-Score	Soporte
0	0.50	0.40	0.44	10
1	0.86	0.60	0.71	10
2	0.33	0.40	0.36	10
3	0.31	0.40	0.35	10
4	0.67	0.80	0.73	10
5	0.75	0.30	0.43	10
6	0.67	0.80	0.73	10
7	0.41	0.90	0.56	10
8	0.80	0.40	0.53	10
9	1.00	0.50	0.67	10
Exactitud			0.55	100
Promedio Macro	0.95	0.95	0.55	100
Peso promedio	0.95	0.95	0.55	100

Tabla 9. Reporte de clasificación conjunto de grabaciones propias CNN Espectrogramas.

	Precisión	Recall	F1-Score	Soporte
0	0.50	0.40	0.44	10
1	0.86	0.60	0.71	10
2	0.33	0.40	0.36	10
3	0.31	0.40	0.35	10
4	0.67	0.80	0.73	10
5	0.75	0.30	0.43	10
6	0.67	0.80	0.73	10
7	0.41	0.90	0.56	10
8	0.80	0.40	0.53	10
9	1.00	0.50	0.67	10
Exactitud			0.55	100
Promedio Macro	0.95	0.95	0.55	100
Peso promedio	0.95	0.95	0.55	100

La Mecatrónica en México, Noviembre 2021, Vol. 9, No. 1, páginas 64 – 84
Disponible en línea en www.mecamex.net/revistas/LMEM
ISSN: 2448-7031, Asociación Mexicana de Mecatrónica A.C

6. Discusión de resultados

Tras haber realizado las pruebas con ambos modelos, podemos realizar ahora una comparativa entre los resultados obtenidos. En la Tabla 10 se puede observar la comparación entre el desempeño de ambos modelos. La CNN señales entrenó mas lentamente que la CNN espectrogramas, sin embargo podemos observar que las gráficas de entrenamiento fueron más estables que las de CNN espectrogramas. Por otro lado la exactitud de la CNN espectrogramas fue mucho mayor en un tiempo de entrenamiento que disminuye casi por mitad que la CNN señales.

Tabla 10. Comparativa de resultados conjunto de pruebas FSDD.

	Tiempo de entranamiento	Precisión	Recall	F1-Score	Soporte
CNN Señales	33 minutos	87.6%	87.6%	87.6%	87.6%
CNN Espectrogramas	14 minutos	94.7%	94.7%	94.7%	94.7%

Por otro lado, encontramos que el desempeño de ambas redes varia de manera significativa al validarlos con el conjunto de grabaciones propias (Tabla 11), ninguno de los dos alcanzó exactitudes similares a las pruebas realizadas con la base de datos FSDD. Aquí existen muchos factores que pueden explicar este comportamiento, algunos de ellos son las condiciones de grabación, la correcta pronunciación de los dígitos, ruido de fondo, frecuencia muestral, entre otras. Sin embargo, con estas pruebas podemos observar que la CNN señales obtuvo una exactitud mas alta que la CNN espectrogramas a diferencia con las pruebas de la base de datos FSDD, dónde la CNN espectrogramas obtuvo un mejor desempeño que la CNN señales.

Tabla 11. Comparativa de resultados conjunto de grabaciones propias.

	Precisión	Recall	F1-Score	Soporte
CNN Señales	55.0%	55.0%	55.0%	55.0%
CNN Espectrogramas	49.0%	49.0%	49.0%	49.0%

7. Conclusiones

En ambos casos, tanto con los datos de la señal como en los espectrogramas, se puede apreciar que uno de los factores mas importantes que pueden influir en los resultados de clasificación son las condiciones de grabación como la calidad del audio, el ruido, la pronunciación, etc. Sin embargo podemos deducir por los resultados que el uso de espectrogramas es mas eficiente que el uso de los datos de la señal, con ellos se puede llegar a una mejor exactitud siempre y cuando las condiciones de grabación y la calidad de los datos sea la misma, esto podemos apreciarlo con la exactitud del 94.7 % obtenida con el conjunto de datos de prueba de la base de datos *FSDD*. Sin embargo, con el uso de grabaciones propias podemos ver una baja considerablemente grande en la evaluación de ambos modelos.

De aquí podemos deducir que la *CNN señales* posee una sensibilidad mayor a la *CNN espectrogramas* por lo que en condiciones de grabación distintas es mas confiable el uso de las señales de audio, esto se debe al momento de convertir las grabaciones a imágenes de espectrograma, las condiciones de grabaciones influyen mucho en la extracción de características. El

uso de imágenes de espectrograma puede ofrecer mejores resultados mas confiables al momento de una clasificación.

Algunas consideraciones a tomar en cuenta para mejorar los resultados tanto de la *CNN señales* como de la *CNN espectrogramas*, además del ajuste de hiperparámetros, podría ser incluir un mayor número de grabaciones propias en el conjunto de entrenamiento, dónde ambas redes podrían aprender a identificar los patrones de las señales y las imágenes con diferentes condiciones y calidad de grabación.

Otra consideración importante podría ser identificar el ruido en las grabaciones y a tras cierta aparición de ruido aplicar algún tipo de filtrado para mejorar la calidad de las grabaciones cuidando no perder información importante de la señal.

En resumen se puede concluir que el uso de CNN para la clasificación de señales puede generalizar mejor e identificar características importantes de la señal al momento de clasificar.
Las imágenes de espectrograma son excelentes para el uso de CNN de clasificación de señales de audio, pero es extremadamente sensible a la aparición de ruido y calidad de grabación.

La aparición de ruido, la pronunciación y la calidad de la grabación son los factores mas importantes que intervienen al momento del entrenamiento del modelo, afectan radicalmente los resultados.

Referencias

[1] Jackson, Z. Souza, C. Flaks, J. Pan, Y. Nicolas, H. & Thite, A. (2018). Free Spoken Digit Dataset. https://github.com/Jakobovski/free-spoken-digit-dataset
[2] LeCun, Y. & Cortes, C. (2010). MNIST handwritten digit database.
[3] Emre C¸ akir, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen, “Convolutional recurrent neural networks for polyphonic sound event detection,” CoRR, vol. abs/1702.06286, 2017.
[4] Jinkyu Lee and Ivan Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in Interspeech 2015. September 2015, ISCA - International Speech Communication Association.
[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” CoRR, vol. abs/1409.0473, 2014.
[6] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” CoRR, vol. abs/1502.03044, 2015.
[7] Yong Xu, Qiuqiang Kong, Qiang Huang, Wenwu Wang, and Mark D. Plumbley, “Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging,” CoRR, vol. abs/1703.06052, 2017.
[8] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D. Plumbley, “Large-scale weakly supervised audio classification using gated convolutional neural network,” CoRR, vol. abs/1710.00343, 2017.
[9] Hossain, M. A., & Alam Sajib, M. S. (2019). Classification of Image using Convolutional Neural Network (CNN). *Global Journal of Computer Science and Technology*, *19*(2), 13–18. https://doi.org/10.34257/gjcstdvol19is2pg1
[10] da Silva, B., Happi, A. W., Braeken, A., & Touhafi, A. (2019). Evaluation of classical Machine Learning techniques towards urban sound recognition on embedded systems. *Applied Sciences (Switzerland)*, *9*(18). https://doi.org/10.3390/app9183885
[11] Virtanen, T., Plumbley, M. D., & Ellis, D. (2017). *Computational analysis of sound scenes and events*. *Computational Analysis of Sound Scenes and Events* (pp. 1– 422). Springer International Publishing. https://doi.org/10.1007/978-3-319-63450-0

[12] Stein, J. (2000). *Digital Signal Processing. A Computer Science Perspective.* A Wiley-Interscience Publication, John Wiley & Sons Inc.
[13] Petetin, Y., Laroche, C., & Mayoue, A. (2015). Deep neural networks for audio scene recognition. *2015 23rd European Signal Processing Conference, EUSIPCO 2015*, 125–129. https://doi.org/10.1109/EUSIPCO.2015.7362358
[14] Cakir, E., Heittola, T., Huttunen, H., & Virtanen, T. (2015). Polyphonic sound event detection using multi label deep neural networks. *Proceedings of the International Joint Conference on Neural Networks*, *2015-Septe*. https://doi.org/10.1109/IJCNN.2015.7280624
[15] Abeßer, J. (2020). A review of deep learning based methods for acoustic scene classification. *Applied Sciences (Switzerland)*, *10*(6). https://doi.org/10.3390/app10062020
[16] Giobergia, F. Free Spoken Digit Dataset Classification Problem.
[17] Nasr, S., Quwaider, M., & Qureshi, R. (2021, July). Text-independent Speaker Recognition using Deep Neural Networks. In *2021 International Conference on Information Technology (ICIT)* (pp. 517-521). IEEE.
[18] Sharmin, R., Rahut, S. K., & Huq, M. R. (2020). Bengali spoken digit classification: A deep learning approach using convolutional neural network. *Procedia Computer Science*, *171*, 1381-1388.
[19] Das, S., Yasmin, M., Arefin, M., Taher, K. A., Uddin, M. N., & Rahman, M. A. (2021, July). Mixed bangla-english spoken digit classification using convolutional neural network. In *International Conference on Applied Intelligence and Informatics* (pp. 371-383). Springer, Cham.
[20] Rao, K. S., & Manjunath, K. E. (2017). *Speech recognition using articulatory and excitation source features*. Springer.
[21] J.W. Picone, Signal modeling techniques in speech recognition. Proc. IEEE 81, 1215–1247 (1993)
[22] L. Rabiner, B.-H. Juang, B. Yegnanarayana, Fundamentals of Speech Recognition (Pearson Education, London, 2008)
[23] S. Furui, Comparison of speaker recognition methods using statistical features and dynamic features. IEEE Trans. Acoust. Speech Sig. Proc. 29, 342–350 (1981)
[24] J.S. Mason, X. Zhang, Velocity

Biografía de Autores

Hernández Montejano Carlos Oliver. Ingeniero en sistemas computacionales por el Tecnológico Nacional de México campus La Piedad. Durante dos años trabajó como desarrollador de software para la empresa Glouu Technologies, participando en proyectos empresariales orientados al sistema SAP. Por los siguientes tres años participó como gerente de servicio de soporte empresarial en la misma empresa como director de proyectos de desarrollo de software, consultoría SAP, atención a cliente y definición de procesos internos de soporte correctivo y preventivo de software. Actualmente estudiante de Maestría en Ciencias en Inteligencia Artificial en la Universidad Autónoma de Querétaro, su area de interés es el estudio y procesamiento de Señales Biomédicas.

González Huerta Rodrigo. Estudió la Licenciatura en Tecnología en la Universidad Nacional Autónoma de México. Actualmente es estudiante de la Maestría en Ciencias en Inteligencia Artificial en la Universidad Autónoma de Querétaro con interés en el estudio de Señales de Audio.

Tovar Arriaga Saúl. Es profesor de tiempo completo en la Universidad Autónoma de Querétaro desde 2010 donde imparte cursos en la Facultad de Ingeniería a nivel licenciatura y posgrado. Obtuvo su grado de Doctor en Ciencias Biomédicas (Dr.rer.hum.biol.) en la Universidad de Erlangen-Nuremberg, Alemania, el grado de Maestría en Ciencias en Mecatrónica (M.Sc.) en la universidad de Siegen, Alemania, y título de Ingeniero en Electrónica en el Instituto Tecnológico de Querétaro. Es miembro del Sistema Nacional de Investigadores nivel 1 y perfil PRODEP. Ha sido encargado y colaborador en diferentes proyectos de investigación con fondos del Ministerio Federal de Educación e Investigación Alemania (BMFB), INNOVAPYME, FOMIX y PROMEP, relacionados a navegación de instrumentos

quirúrgicos, procesamiento de imágenes y software embebido. Ha impartido conferencias a nivel nacional e internacional, publicado varios artículos indizados en el JCR en el área de cirugía asistida por computadora, robótica y física médica. Es actualmente coordinador de la Maestría en Ciencias en Inteligencia Artificial, Presidente del Capítulo de Computación Inteligente del IEEE de la sección Querétaro y es representante del Cuerpo Académico de Optimización y Computación Avanzada.