ISSN 1870-4069

Reconocimiento del habla en videos digitales usando redes neuronales convolucionales

Cesar E. Embriz-Islas, Cesar Benavides-Alvarez,
Carlos Avilés-Cruz, Arturo Zúñiga-López

Universidad Autónoma Metropolitana,
Unidad Azcapotzalco,
Departamento de Electrónica,
México

{al2211800610, cesarbenavides,
caviles, azl}@azc.uam.mx

Resumen. El reconocimiento del habla con contexto visual es una técnica que utiliza el procesamiento digital de imágenes, para detectar el movimiento de los labios dentro de los cuadros de imagen de un video para predecir las palabras que está exclamando un hablante. Aunque ya existen modelos con resultados sobresalientes, la mayoría están enfocados en ambientes muy controlados con pocas interacciones del hablante. En este trabajo, se propone una nueva implementación de un modelo, basado en redes neuronales convolucionales (CNN), considerando los cuadros de imagen y tres modelos de uso del audio, por medio de espectrogramas. Los resultados obtenidos son muy alentadores en el campo del reconocimiento automático del habla.

Palabras clave: CNN, inteligencia artificial, aprendizaje profundo, reonocimiento del habla.

Speech Recognition in Digital Videos Using Convolutional Neural Networks

Abstract. Visual contextual speech recognition is a technique that uses digital image processing to detect lip movement within video image frames to predict the words a speaker is exclaiming. Although there are already models with outstanding results, most are focused on highly controlled environments with few speaker interactions. In this work, a new implementation of a model is proposed, based on convolutional neural networks (CNN), considering image frames and three models of audio use, by means of spectrograms. The results obtained are very encouraging in the field of automatic speech recognition.

Keywords: CNN, artificial intelligence, deep learning, speech recognition.

1. Introducción

El reconocimiento del habla con contexto visual es el proceso de detectar las palabras o vocablos emitidos por una persona a través de un video. Este tema, en los últimos años se ha desarrollado con gran interés, por las múltiples aplicaciones que puede tener, algunas de ellas son: el reconocimiento del habla para videos de seguridad, lectura de labios para oyentes con problemas de audición, detección de videos con audio alterado, entre otros.

Para reconocer el habla, en los estudios de Assael et al. [1] y de Garg et al. [6] se basan en procesar el video digital como cuadros de imagen (*frames*[1], nombre en en inglés), para extraer las regiones de los labios que funcionan como entrada para un modelo basado en redes neuronales convolucionales (*CNN*). Por otro lado, los trabajos [5, 2, 8] han demostrado que tomar en cuenta el audio (además de las regiones de los labios), mejoran considerablemente la precisión del reconocimiento.

En el presente artículo, se propone un nuevo modelo de inteligencia artificial basado en redes neuronales convolucionale para trabajar con videos de personas para reconocer el habla de forma automática. El modelo propuesto es un híbrido de las arquitecturas que se han utilizado, y considerando una forma diferente para entrenar los videos, se contempla la utilización de las palabras cortas y no las oraciones completas.

En la figura 1, se describen los bloques principales de nuestra propuesta para el reconocimiento automático del habla. Inicialmente se parte de una base de datos de videos (figura 1a); posteriormente, se pre-procesan los videos para extraer cuadros de imagen de la región de interés (labios), y asimismo, se procesa el audio teniendo una representación en imagen (figura 1b). Finalmente, se extraen las características importantes de las imágenes procesadas, para obtener la clasificación o reconocimiento de la palabra, ver figura 1c.

La parte del pre-procesamiento, comprende la delimitación de los labios del hablante, por cada cuadro de imagen extraído del video. Asimismo, se procesa el audio como un canal independiente considerando tres diferentes tipos de espectrogramas, en donde la elección del tipo de espectrograma adecuado es fundamental para obtener un mejor desempeño en la clasificación.

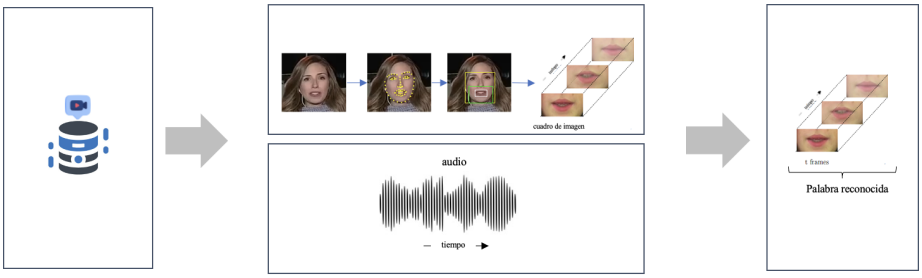
Finalmente, en la parte de clasificación, comprende todo el entrenamiento para asociar el movimiento de los labios con la palabra correcta. Se utiliza una *CNN*, como entrada, se consideran los cuadros de imagen de los labios, previamente procesados.

De igual forma para el audio, es una entrada para otra *CNN* independiente a la de los labios. Estos procesos, sirven para extraer las características de ambas entradas descritas, y con una capa *softmax*[2] se clasifican las palabras por el movimiento de los labios; más detalles serán descritos en el capítulo de Metodología.

El resto del artículo esta constituido de la Metodología, descrita en el capítulo 3. En el capítulo 4 se describe la experimentación y resultados; finalmente, las conclusiones y perspectivas son descritas en el capítulo 5.

[1] Mínima imagen completa registrable de un video.

[2] La función softmax convierte un vector de K números reales en una distribución de K resultados posibles.



(a) Base de datos (b) Pre-procesamiento (c) Clasificación

Fig. 1. Esquema general del proceso de reconocimiento del habla.

2. Estado del arte

El artículo de Belhan et al. [2] hace un comparativo entre modelos de reconocimiento del habla inglesa, considerando un entrenamiento sin audio implementando la red LipNet, otro con audio (LipNet-audio) y finalmente, un tercero con el audio aplicando la transformada de Fourier de tiempo corto, *STFT* por sus siglas en inglés, (LipNet-audio-STFT).

El modelo LipNet desarrollado por Assael et al. [1] es clásico y conocido; se apoya de capas *GRU* bidireccionales en lugar de unidireccionales. Por otro lado LipNet-audio procesa los datos visuales utilizando capas GRU unidireccionales y los datos de audio como señales 1D sin STFT. Finalmente LipNet-audio-STFT, procesa los datos visuales igual que LipNet-audio y los datos de audio con señales 3D con STFT con una redimensión a resolución 64×64 píxeles.

Al utilizar datos de audio, el modelo mejora considerablemente los resultados. Por otro lado, el modelo se desempeña mejor cuando se utiliza el audio con señales 1D en comparación con 3D y STFT, sin considerar el tiempo computacional extra por aplicar STFT. El considerar una resolución completa del audio con STFT mejoraría los resultados pero con más costo computacional.

Por otro lado, el trabajo de Feng [5] propone un modelo con una red neuronal recurrente multimodal, *m-RNN* (por sus siglas en inglés), para el reconocimiento del habla audiovisual. La estructura del modelo, consta de dos componentes: una parte visual y una de audio.

La parte visual, contiene una capa CNN más una *LSTM* bidireccional; y la parte de audio, una LSTM bidireccional. Ambas partes contienen capas de estado ponderado, para generar resultados semánticamente coherentes para la fusión. Basado en las capas de estado ponderado, la RNN multimodal utiliza una capa multimodal para fusionar ambas modalidades, y una capa *softmax* para la salida.

De los estudios más recientes, Jeon et al. [8] presenta una arquitectura de lectura de labios para el reconocimiento del habla visual a nivel de oración. La arquitectura está compuesta por tres módulos de extracción de características visuales, y se aplican múltiples métodos de extracción de características visuales, que logran una mejor predicción del movimiento de los labios. En los estudios anteriormente mencionados, se ha utilizado la base de datos GRID [4].

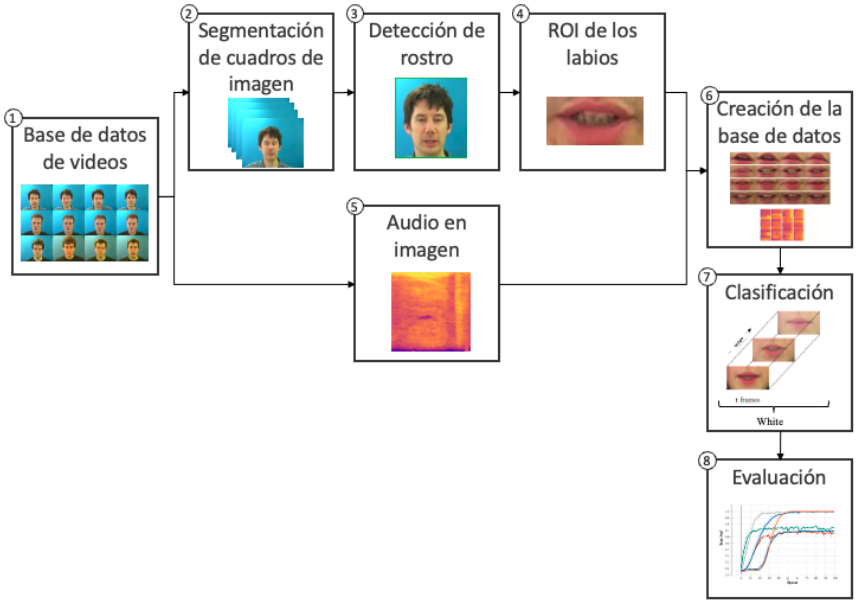


Fig. 2. Metodología del sistema de reconocimiento del habla.

Otro enfoque importante es el que propone Rossler et al. [7] utilizando la base de datos [9], la cual está constituida por falsificaciones faciales, plantea sobre la construcción de un modelo que reconoce falsificación de videos cuando el audio no coincide con el movimiento de labios.

Para detectar el video alterado, propone un extractor espacio-temporal de características, seguido de una CNN temporal para la lectura de labios. Posteriormente, se congela el extractor de características y se ajusta la red temporal de falsificación de datos.

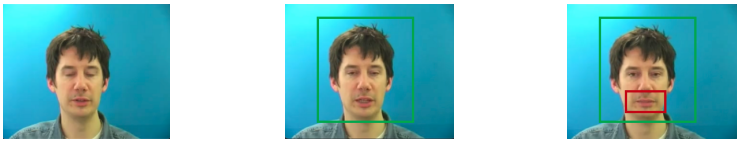
3. Metodología

El modelo propuesto y los bloques metodológicos se muestra en la figura 2, cada uno de los ocho bloques constitutivos se describen a continuación:

3.1. Base de datos de videos

Se crea a partir de las palabras extraídas de la base de datos *GRID* [4], el corpus de oraciones audiovisuales de múltiples hablantes. El corpus, consta de grabaciones de audio y video, donde se aprecia perfectamente el rostro. La resolución del video es de 360×288 píxeles y un audio a $50kHz$.

El corpus, contiene 1,000 oraciones pronunciadas por cada uno de los 33 hablantes (17 hombres y 16 mujeres), con un total de 33,000 oraciones. La estructura de las oraciones es la siguiente:



(a) Cuadro extraído (b) Detección del rostro (c) Detección de ROI

Fig. 3. Detección del rostro y ROI de los labios.

$$\text{comando}(4) + \text{color}(4) + \text{preposición}(4) + \text{letra}(25) + \text{dígito}(10) + \text{adverbio}(4)$$

donde (#) indica la cardinalidad del conjunto de palabras para cada una de las 6 categorías:

- *comando:* {bin, lay, place, set},
- *color:* {blue, green, red, white},
- *preposición:* {at, by, in, with},
- *letra:* {A,. . . ,Z}/{W},
- *dígito:* {zero,. . . , nueve},
- *adverbio:* {again, now, please, soon}.

Y un ejemplo es: *'set blue by A four please'*.
La base de datos contiene un total de 51 palabras diferentes.

3.2. Segmentación de cuadros de imagen

Se procesan los videos digitales, para extraer 30 cuadros de imagen por segundo del video. En la figura 3a se muestra un cuadro extraído del video.

3.3. Detección del rostro

Por cada cuadro de imagen extraído del video, se detecta el rostro y obtienen las coordenadas que lo comprenden. En la figura 3b se presenta un ejemplo de la detección del rostro dentro del cuadro de color verde.

3.4. Región de interés (*ROI*) de los labios

Una vez detectado el rostro, se obtienen las coordenadas de los labios. Estas últimas coordenadas se utilizan para considerar un área rectangular, y delimitar la *ROI* que se utiliza para extracción de características. Del largo de los labios, se consideran 5 píxeles adicionales, tanto para la izquierda como la derecha; y para la altura, se consideran 10 píxeles, para arriba y abajo; con esto se delimita el rectángulo mencionado. En la figura 3c se muestra un ejemplo de la detección de *ROI* de los labios en color rojo.

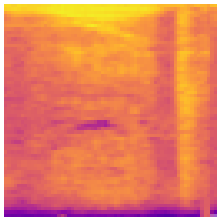


Fig. 4. Espectrograma tipo 1 del audio correspondiente a la palabra "*white*".

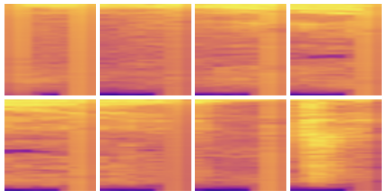


Fig. 5. Espectrograma tipo 2 del audio correspondiente a la palabra "*white*".

3.5. Transformación audio en imagen

Por cada segmento de audio (señal unidimensional), se realiza una transformación Tiempo-Frecuencia (señal bi-dimensional, imagen), llamado espectrograma, para convertir el audio a imagen; la transformación se aplica con la ecuación 1:

$$V_g f(x,\omega) = \int_{-\infty}^{\infty} f(t) g(t-x) e^{-2\pi i t\dot{\omega}}\, dt, \tag{1}$$

donde ω representa el tiempo del audio, y para este proyecto, se contemplan tres posibles variaciones.

1. Un único espectrograma de la palabra, ver Figura 4. Se toma el audio completo de la palabra para realizar la transformación Tiempo-Frecuencia.
2. En la figura 5, se observa la cantidad de espectrogramas alineados al mismo número de cuadros de imagen del video. En este caso, se genera un espectrograma para cada *frame*. Así, se tienen igual número de *frame* y espectrogramas.
3. En este última opción, cada segundo de la palabra se divide en 10 intervalos equi-espaciados, obteniendo solo diez espectrogramas por segundo (ver figura 6).

Este preprocesamiento se ejecuta en paralelo al de la extracción de la ROI de los labios.

3.6. Creación de base de datos

Con los pre-procesamientos anteriores, se genera una nueva base de datos, que comprende 30 cuadros por segundo de la ROI de los labios, con dimensiones de 160×80. En la figura 7 se muestran ejemplos de cuadros de imagen procesados con la ROI seleccionada.

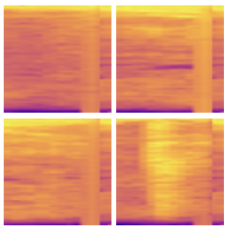


Fig. 6. Espectrograma tipo 3 del audio correspondiente a la palabra "*white*".

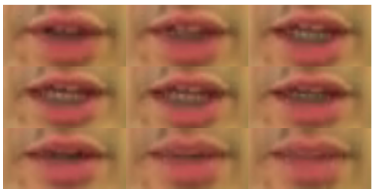


Fig. 7. Labios procesados del cuadro de imagen correspondientes a la palabra "*again*".

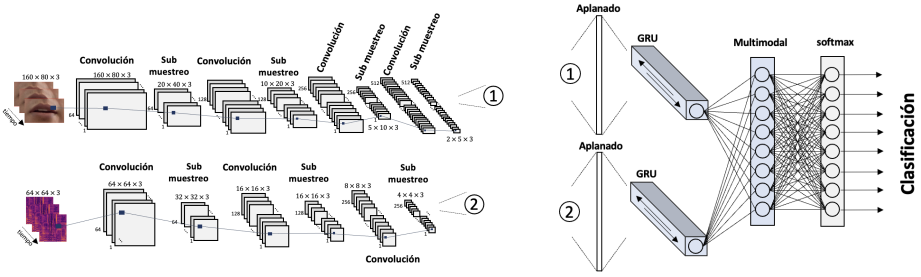
Por otro lado, también se tienen imágenes con un tamaño de 64×64 pixeles para los diferentes tipos de espectrogramas planteados anteriormente vistos en las figuras 4, 5 y 6.

3.7. Clasificación - Modelo de aprendizaje profundo -

Para la clasificación se plantea un modelo de Inteligencia Artificial basado en aprendizaje profundo. Particularmente, se usan redes neuronales convolucionales, redes neuronales retroalimentadas y redes neuronales de clasificación.

El modelo propuesto empata la información del movimiento de los labios, con el espectrograma proveniente del audio. El modelo de aprendizaje profundo propuesto (ver figura 8a) es detallado a continuación:

(a) Una CNN para la extracción de características importantes de las imágenes de los labios. La CNN consta de cuatro capas convolucionales; la primera capa la constituyen 64 filtros con un kernel de $(3, 5, 5)$, un paso de $(1, 2, 2)$, y un relleno igual; para la segunda, 128 filtros con un kernel de $(3, 5, 5)$; la tercera y cuarta tienen 256 y 512 filtros, respectivamente, y ambas utilizando los parámetros por defecto con un kernel de $(3, 3, 3)$. Cada una de las capas utiliza una función de activación *relu* y cada capa es acompañada por un sub-muestreo (*max-pooling*) y una normalización (*batch_normalization*). Asimismo, la entrada de esta red, es una serie de cuadros de imagen de $160 \times 80 \times 3$ píxeles que representan la palabra expresada.
(b) Una *CNN* para la extracción de características importantes del espectrograma. La *CNN* consta de tres capas convolucionales, tres de sub-muestreo y tres de normalización. Esta red, es una copia hasta la tercer capa de la anteriormente descrita, la diferencia es que existen tres entradas diferentes, que hacen referencia a los tres tipos de espectrogramas.



(a) Primera parte de la red

(b) Segunda parte de la red

Fig. 8. Modelo propuesto.

(c) Dos capas de *aplanamiento* sobre tiempo distribuido, una por cada *CNN*.
(d) Dos capas bidireccionales *GRU* por cada *aplanamiento*. Para ambas redes, los filtros que se aplican son 2048 y 1024.
(e) Una capa densa por cada red, cada una de 512 filtros con activación *leakyrelu*.
(f) Una capa *multimodal* que concatene las dos salidas anteriores.
(g) Una serie de 3 capas densas con filtros descendientes de 512, 256 y 128; todas con una activación *relu*.
(h) Finalmente, una capa *softmax* con 51 clases (número de palabras) para la clasificación.

En la figura 8 se muestra, de manera gráfica, la implementación del modelo, en donde se describen las convoluciones con la reducción de dimensiones de las imágenes y el incremento de filtros; finalmente la red recurrente con la multimodal para definir el clasificador. Por otra parte, en la figura 9 se muestra la implementación del modelo con tensorflow con mucho más detalle de la evolución de dimensiones y la capas que se han aplicado.

4. Experimentación y resultados

Se hace una selección de dos observaciones por palabra y por hablante. Obteniendo un total de 102 videos de palabras por cada hablante, en total 3,366 videos de las 51 clases de palabras. Siguiendo los pasos de la metodología; para los videos, se segmentan en cuadros de imagen, detecta el rostro y se extrae la ROI de los labios; para el audio, se generan los tres tipos de espectrogramas.

En ambos procesos, se guardan las imágenes correspondientes, creando una nueva base de datos. Con el objetivo de tener más estabilidad en el modelo, se genera un aumento de datos, para finalmente entrenar y evaluar el modelo propuesto.

4.1. Procesamiento de los labios

El número de imágenes que genera cada palabra depende de la complejidad, tamaño y pronunciación. Hay algunas palabras cortas como la letra *I*, que la representan seis cuadros de imagen.

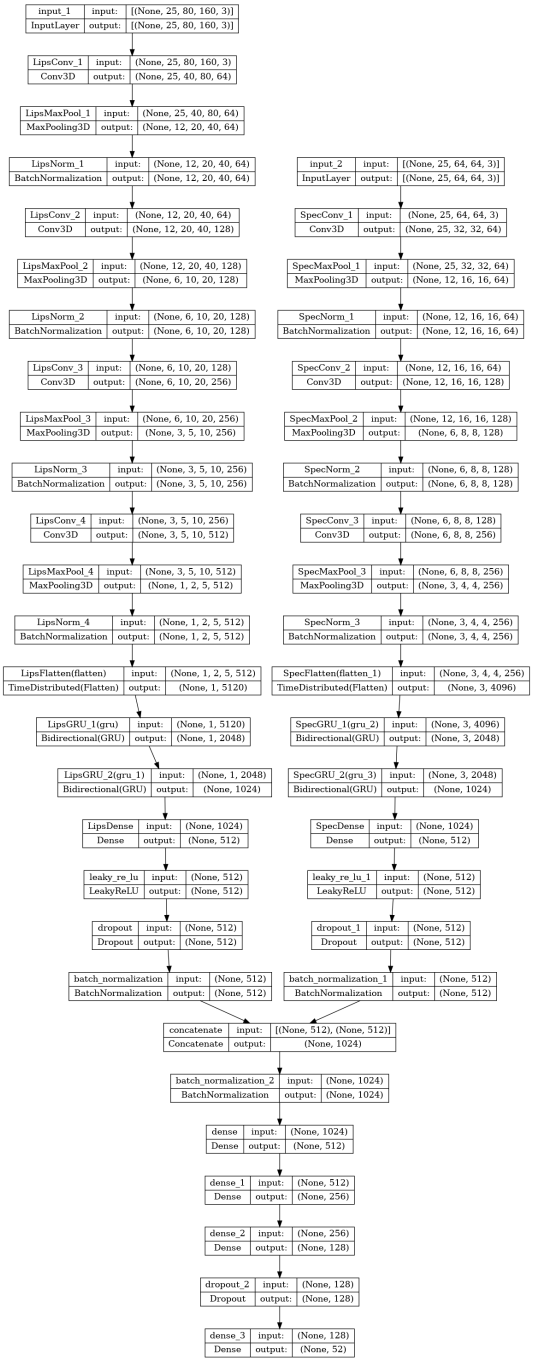


Fig. 9. Modelo implementado en Tensorflow.

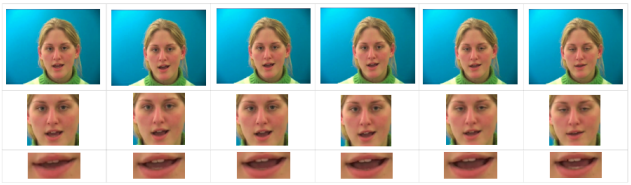


Fig. 10. Procesamiento de la letra *I*.

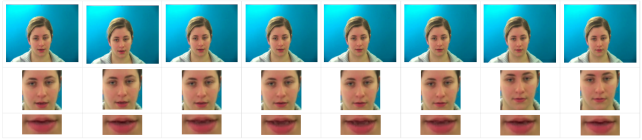


Fig. 11. Procesamiento de la palabra *again*.

En la figura 10 se muestra la etapa del procesamiento; primero, la conversión del video a cuadros de imagen; el segundo, en donde se detecta el rostro por cada cuadro de imagen; y finalmente los labios, que es la región de interés, y entrada más importante del modelo.

En las figuras 11 y 12 se ven dos ejemplos más del procesamiento que se realiza sobre los videos. Para la primera imagen, la componen un total de 8 cuadros de imagen; y para la segunda, un total de 6, similar al de la letra *I*. Con estos ejemplos se ilustra que los cuadros que representan la pronunciación de una letra o palabra no son de un tamaño en específico.

Otro factor importante, es la entonación sobre la palabra que hace cada persona. En la figura 13 se muestra el procesamiento para la palabra *by* de un hablante diferente a la expresada en la figura 12. En este segundo ejemplo, los cuadros de imagen que representan a la palabra son 4, a diferencia de la primera, que muestran 6.

4.2. Procesamiento del audio

Por otro lado, tomando de referencia la palabra *again*, se aplica el procesamiento para el audio, el cual genera tres tipos de secuencias diferentes:

1. Espectrograma tipo 1: esta transformación considera un único espectrograma de toda la palabra sin importar el número de cuadros de imagen que la componen. En la figura 14a se muestra la representación de la palabra.
2. Espectrograma tipo 2: para este caso, se obtienen un número de espectrogramas igual a el número de cuadros de imagen que representa la palabra, es decir, para la palabra *again* del ejemplo anterior, tiene un total de 6 cuadros de imagen, y cada cuadro está representado por un espectrograma como se muestra en la figura 14c.
3. Espectrograma tipo 3: en este tipo hace referencia a la duración de la pronunciación de la palabra, recordando que son 10 espectrogramas por segundo, para la palabra *again* sólo se consideran 4 espectrogramas, que es la parte proporcional del segundo, como se muestra en la figura 14c.

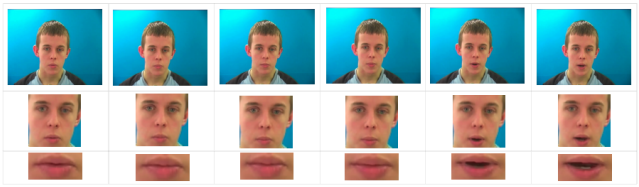


Fig. 12. Procesamiento de la palabra *by*.

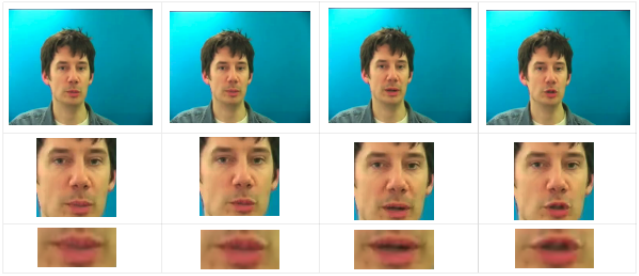


Fig. 13. Procesamiento de la palabra *by*.

De la mima manera, se genera un *aumento de datos*[3][10].

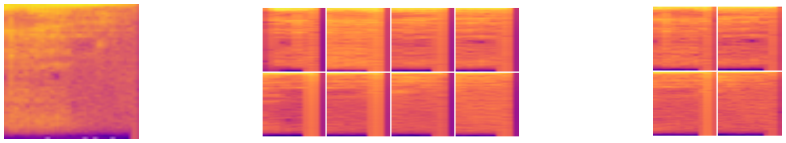
Una vez generada la base de datos aumentada, se divide en dos segmentos, el primero para entrenamiento y el segundo para validación. Para la creación de la partición de entrenamiento $train$, se considera un $77\,\%$ del total de los datos; y para la validación val, se toma el $33\,\%$ restante.

4.3. Entrenamiento

Una vez listo el pre-procesamiento de los datos, se aplica el entrenamiento del modelo de aprendizaje profundo, considerando el conjunto de datos previamente particionado. Se realizan múltiples iteraciones ajustando diferentes parámetros como la función de pérdida (*loss function*), tasa de aprendizaje (*learning rate*), épocas (*epochs*), tamaño de lote (*batch size*), entre otros. Lo siguiente es considerado en la ejecuciones de entrenamiento:

1. Función de pérdida: *entropía cruzada* entre las etiquetas y las predicciones. Se utiliza esta función ya que existen más de dos clases.
2. Optimizador: $Adam$. La optimización de Adam es un método de descenso de gradiente estocástico que se basa en la estimación adaptativa de momentos de primer y segundo orden.
3. Métricas de seguimiento: exactitud (accuracy) y exactitud en la validación (val_accuracy).

[3] El aumento de datos es una técnica utilizada en el aprendizaje automático y la visión por computadora para aumentar el tamaño y la variabilidad de un conjunto de datos de entrenamiento mediante la generación de nuevos ejemplos de entrenamiento a través de varias transformaciones de los datos originales.



(a) Espectrograma tipo 1. (b) Espectrograma tipo 2. (c) Espectrograma tipo 3.

Fig. 14. Espectrogramas de la palabra *again*.

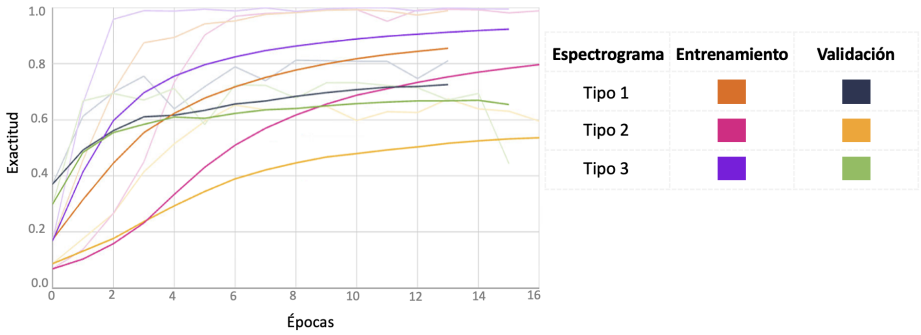


Fig. 15. Exactitud por época, evaluando cada tipo de espectrograma.

4.4. Evaluación

La aplicación del modelo se utiliza el siguiente *hardware*: Una computadora marca *Alienware*, con procesador intel core i9 12th generación, tarjeta gráfica NVIDIA GeForce RTX 3090, 24 GB GDDR6X, memoria RAM DDR5 de 64 GB y 2Tb de SSD.

Por otro lado, las consideraciones para el *software*, son las siguientes: sistema operativo *Ubuntu*, versión 20.04, lenguaje de programación *Python* 3.9.13, librería *OpenCV* [3], librería *face_recognition*[4] 1.2, librería *tensorflow*[5] 2.r2.9

Con las especificaciones anteriores, y en la evaluación, se consideran tres ejecuciones del modelo, una por cada tipo de espectrograma planteado. En la figura 15 se muestran las tres diferentes ejecuciones que se han realizado, puntualizando la exactitud por cada época para los conjuntos de entrenamiento y validación.

Los resultados obtenidos, tabla 1, muestran que los tres tipos de espectrograma presentan un comportamiento similar de aprendizaje. Sin embargo, el tipo 1 es el que demuestra un mejor desempeño con una alta exactitud tanto en entrenamiento como en validación, y con un bajo número de épocas.

Cabe destacar que el tipo 1 corresponde a una única imagen para el audio transformado. En contraste, el tipo 2 y 3 presentan una exactitud más baja en ambas fases de evaluación.

[4] https://github.com/ageitgey/face_recognition\#face-recognition
[5] https://www.tensorflow.org/?hl=es-419

Tabla 1. Exactitud por tipo de espectrograma.

Tipo	Entrenamiento (exactitud)	Validación (exactitud)
1	0.85	0.73
2	0.78	0.53
3	0.92	0.65

5. Conclusiones y trabajo futuro

Se pueden plantear las siguientes conclusiones, se analizó y se programó exitosamente el modelo de aprendizaje profundo propuesto. Por otro lado, se evaluaron las tres representaciones Tiempo-Frecuencia de audio en espectrogramas correspondientes (1) Un espectrograma por palabra, (2) un espectrograma por cuadro (frame), y (3) diez espectrogramas por segundo.

Los resultados obtenidos indican que los tres tipos de espectrogramas muestran un comportamiento similar de aprendizaje, pero el tipo 1, que corresponde a una única imagen para el audio transformado, muestra el mejor desempeño con una buena exactitud tanto en el conjunto de entrenamiento como en el conjunto de validación.

Los resultados sugieren que el tipo 1 es una buena opción para la red propuesta. Asimismo, es importante resaltar que al utilizar el tipo 1 del espectrograma, el tiempo de transformación, en la etapa del pre-procesamiento, es más eficiente.

Hasta el momento no se ha encontrado un trabajo que utilice la base de datos *GRID* [4] como en éste. Los videos utilizados, han sido descompuestos para tener la palabra, letra o número de manera particular y no dentro de una oración, como en el video original. Debido a lo anterior, es complicado tener una referencia para medir la eficiencia del modelo.

Como trabajo futuro, se propone explorar otras arquitecturas y optimización de hiper-parámetros para mejorar la exactitud en clasificación. También sería interesante evaluar el rendimiento del modelo en otros conjuntos de datos e incluso en videos en idioma español.

Referencias

1. Assael, Y. M., Shillingford, B., Whiteson, S., de Freitas, N.: LipNet: End-to-end sentence-level lipreading (2017) https://openreview.net/forum?id=BkjLkSqxg
2. Belhan, C., Fikirdanis, D., Cimen, O., Pasinli, P., Akgun, Z., Yayci, Z. O., Turkan, M.: Audio-visual speech recognition using 3D convolutional neural networks. In: 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 1–5 (2021) doi: 10.1109/ASYU52992.2021.9599016
3. Bradski, G.: The opencv library. Dr. Dobb's Journal: Software Tools for the Professional Programmer, vol. 25, no. 11, pp. 120–123 (2000)
4. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421–2424 (2006) doi: 10.1121/1.2229005

5. Feng, W., Guan, N., Li, Y., Zhang, X., Luo, Z.: Audio visual speech recognition with multimodal recurrent neural networks. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 681–688 (2017) doi: 10.1109/IJCNN.2017.7965918
6. Garg, A., Noyola, J., Bagadia, S.: Lip reading using CNN and LSTM. Technical report, Stanford University (2016)
7. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5039–5049 (2021) doi: 10.1109/CVPR46437.2021.00500
8. Jeon, S., Elsharkawy, A., Kim, M. S.: Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition. Sensors, vol. 22, no. 1, pp. 72 (2021) doi: 10.3390/s22010072
9. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M.: FaceForensics++: Learning to detect manipulated facial images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–11 (2019) doi: 10.1109/iccv.2019.00009
10. Shorten, C., Khoshgoftaar, T. M.: A survey on image data augmentation for deep learning. vol. 6, no. 1, pp. 1–48 (2019)