Detección de Contenido Sensible en Audio y Vídeo mediante Espectrogramas y Aprendizaje por Transferencia

Daniel Povedano Álvarez, Ana Lucila Sandoval Orozco, Luis Javier García Villalba*

Grupo de Análisis, Seguridad y Sistemas (GASS)
Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA)
Facultad de Informática, Despacho 431, Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid
Email: {dpovedano, asandov, javierv}@ucm.es

***Resumen*—Con la creciente proliferación del contenido multimedia en línea, surge la necesidad de garantizar la seguridad digital de los usuarios. La detección de contenido sensible en vídeos representa un desafío crítico para los investigadores y profesionales de la seguridad. Este artículo examina cómo el análisis de audio puede ser una herramienta efectiva para filtrar contenido para adultos sin sacrificar el rendimiento, lo que permite la automatización en la detección de material sensible en dispositivos digitales. Para mejorar esta capacidad, se investigaron métodos de extracción de características acústicas y se evaluaron para la detección de contenido sensible en vídeos. Utilizando una CNN pre-entrenada y utilizando espectrogramas de *log Mel* y aumento de datos, obtuvimos un 86,4 % de *F1-measure* en el conjunto de datos *Pornography-2K*. Finalmente, en la clasificación de vídeos completos se obtuvo un 90,4 % de *F1-measure*.**

***Index Terms*—Aprendizaje profundo, Espectrogramas de Mel, Clasificación de vídeos, Análisis de audio**

Tipo de contribución: *Investigación original (límite 8 páginas)*

I. Introducción

La presencia de contenido sensible en vídeos puede tener consecuencias graves, especialmente en entornos donde los usuarios son vulnerables, como plataformas de redes sociales, aplicaciones de mensajería y entornos de juego en línea. La exposición a este tipo de contenido puede provocar daños psicológicos, traumas emocionales y problemas de seguridad, especialmente para niños y adolescentes. Se ha visto que la exposición temprana a la pornografía y la exposición no regulada/excesiva a la pornografía durante los años formativos de la adolescencia tienen diversos efectos nocivos a largo plazo sobre la maduración sexual, el comportamiento sexual, la adicción a Internet y el desarrollo general de la personalidad [1]. Tres cuartas partes (73 %) de los adolescentes han visto pornografía online a la edad de 17 años, con la edad promedio de la primera exposición a los 12 años, según el informe de *Common Sense Media, "Teens and Pornography"* [2]. Gran parte de la exposición fue por accidente, con el 58 % diciendo que no buscaron los vídeos y fotos sexualmente explícitos, sino que los encontraron mientras navegaban por la web, en las redes sociales o a través de motores de búsqueda o al hacer clic en anuncios.

El método propuesto en este trabajo, destaca por su capacidad para identificar patrones específicos en el audio que puedan indicar la presencia de contenido sensible (sexual), lo cual es especialmente relevante en un contexto donde la tecnología *"Deepfake"* y el uso indebido de Inteligencia Artificial Generativa para crear contenido pornográfico indebido (como por ejemplo pornografía infantil) plantean riesgos significativos.

La utilización de espectrogramas para analizar el audio extraído de vídeos abre nuevas posibilidades en la lucha contra la difusión de contenido inapropiado en plataformas digitales. Este enfoque no solo ofrece una alternativa técnica avanzada para la detección de contenido sexual, sino que también subraya la importancia de desarrollar soluciones tecnológicas que sean capaces de adaptarse y responder a las técnicas cada vez más sofisticadas utilizadas para eludir los sistemas de moderación de contenido existentes.

En este contexto, este trabajo propone explorar las capacidades y limitaciones de los espectrogramas en la detección de contenido sensible en audio mediante Aprendizaje por Transferencia (AT, del inglés *Transfer Learning*),una técnica comúnmente utilizada en el Aprendizaje Profundo (AP, del inglés *Deep Learning*) que permite transferir el conocimiento adquirido en un dominio específico a otro, con el fin de contribuir a un entorno digital más seguro y responsable.

Más específicamente, este trabajo se enfoca en la detección de contenido sensible mediante AT y Redes Neuronales Convolucionales (del inglés *Convolutional Neural Network* (CNN)) empleando el modelo *ConvNeXt V2* pre-entrenado con *ImageNet*. Además, se investigará cómo el uso de un modelo pre-entrenado en un conjunto de datos que no incluye espectrogramas produce un rendimiento excelente en el conjunto de datos *Pornography-2k*, formado por 1000 vídeos pornográficos y 1000 no pornográficos.

El resto del trabajo se organiza como sigue: en la Sección II se presentan brevemente algunos trabajos recientes publicados para la detección de contenido sensible mediante la clasificación de audio, especialmente los enfoques basados en AP. En la Sección III se detalla la propuesta del trabajo. En la Sección IV se describe la configuración de los experimentos realizados. Finalmente, en la Sección V se muestran y discuten los resultados obtenidos y por último, las conclusiones de la investigación se incluyen en la Sección VI.

ISBN:978-84-09-62140-8

II. Trabajos relacionados

Detectar contenido sensible en archivos de audio representa un desafío considerable, y su abordaje demanda el uso de técnicas avanzadas de AP. Este análisis se centra en explorar y examinar las investigaciones más recientes en este ámbito, con especial atención en cómo las metodologías de AP están siendo aplicadas para atenuar este problema. En esta sección se revisa las contribuciones más significativas en el campo de la detección de contenido sensible utilizando características acústicas mediante Aprendizaje Automático (AA) y AP.

Uno de los primeros trabajos que utilizó las características del audio para la detección de contenido sensible utilizando AP fue [3]. Los autores proponen un método de reconocimiento de pornografía en vídeo en vivo *stream* que utiliza características profundas multimodales con atención controlada. Para ello, extraen dichas características (espaciales, de movimiento y de audio) de la transmisión de vídeo en vivo utilizando CNN. En cuanto al conjunto de datos, los autores recolectaron un conjunto de datos reales de Internet, conocido como *BJUT streamer dataset* (BJUTSD), que incluye 1.081 vídeos en directo no pornográficos y 2.511 vídeos porno en directo. Respecto a la implementación técnica, los tres modelos CNN se inicializaron con un modelo pre-entrenado de *ImageNet* y se ajustaron con datos de entrenamiento de vídeos en directo. Además, realizaron aumento de datos (rotación, traslación, etc) sobre los fotogramas (no implementaron ningún aumento de aumento de datos sobre el audio o sobre los espectrogramas). Por último, las características de audio se extrajeron de espectrogramas con una CNN después de convertir la señal de audio de los en espectrogramas mediante una transformación de Fourier. Finalmente, realizaron seis experimentos obteniendo resultados competitivos (74,86 % de exactitud o *accuracy*) demostrando que el método puede reconocer eficazmente los *streams* de pornografía en transmisiones de vídeo en vivo

Song et al. [4] propusieron un enfoque basado en AP con características multimodales, incluyendo descriptores de imagen, características visuales de cada fotograma, características de audio extraídas del vídeo y características de movimiento. Para el detector basado en características de audio, se dividieron los datos de audio en clips de 10 segundos y se extrajeron características utilizando el espectrograma de escala de Mel. Se utilizó un subconjunto aleatorio del conjunto de datos *Pornography-2K* [5] para el entrenamiento y las pruebas SVM, con validación cruzada de 10 veces. Los resultados mostraron una precisión del 88,3 % para el detector basado en textura de movimiento y del 80 % para el detector basado en características de audio.

Song et al. [6] propusieron un esquema de apilamiento multimodal para la detección en línea rápida y precisa de contenidos pornográficos en Internet. Para detectar con precisión el contenido sensible, se extrajeron características visuales y auditivas utilizando un *VGG-16* con una Red Neuronal Recurrente (*Recurrent Neural Network* (RNN)) bidireccional para reflejar los patrones de cambio de señal en el tiempo dentro de cada entrada. El conjunto de datos utilizado fue un subconjunto de *Pornography-2k* y Se recolectó otro conjunto de datos de internet que incluía contenido sensible con aspectos perjudiciales tanto visuales como auditivos. Se seleccionaron aleatoriamente 8.000 instancias de segmentos, de las cuales 5.000 se utilizaron para entrenamiento y 3.000 para pruebas. Cada instancia contenía 4,000 segmentos, tanto de contenido sensible como no sensible. Los resultados experimentales revelaron una precisión del 92,33 %, superando a otros trabajos que emplearon modelos multimodales, así como a los resultados individuales de los clasificadores visual y auditivo (95,33 % y 89,16 % respectivamente).

En el contexto de los mecanismos de atención, a diferencia de las técnicas convencionales que se basan únicamente en características visuales sin tener en cuenta las auditivas, Fu et al. [7] presentaron un sistema de AP unificado denominado *PornNet* que integra subredes duales para la detección de vídeos pornográficos. Para el audio, utilizaron *VGGish*, obteniendo incrustaciones de características de audio, que son espectrogramas *log Mel* y representaciones de imágenes del audio. Posteriormente, para el reconocimiento de las incrustaciones de características de audio, los autores utilizaron la red *RANet*, generando posteriormente resultados de vídeo-audio. El rendimiento de la propuesta se evaluó en un conjunto de datos recopilados por ellos mismos recientemente, mostrando cómo el método propuesto funciona muy bien, alcanzando una exactitud del 93,4 % en el conjunto de datos interno que incluye 1.000 muestras porno junto con 1.000 vídeos normales y 1.000 vídeos *sexys*.

Más recientemente, Lovenia et al. [8] propusieron un enfoque basado centrado únicamente en características acústicas para la detección pornográfica. Este método basado en audio permite filtrar contenidos sensibles explotando diferentes características espectrales. Descubrieron que una CNN entrenada en el espectrograma *log Mel* alcanza el mayor rendimiento en el conjunto de datos *Pornography-800* [9] (que consta de 400 vídeos pornográficos y 400 no pornográficos) que las características MFCC (Coeficientes Cepstrales en las Frecuencias de Mel , del inglés *Mel Frequency Cepstral Coefficients*). Los resultados experimentales mostraron que el espectrograma *log Mel* proporciona mejores características en segmentos de 60 segundos para que los modelos reconozcan los sonidos pornográficos. Por último, para clasificar formas de onda de audio completas en lugar de segmentos, emplearon una técnica de votación de segmento a audio que produce los mejores resultados de audio. Los autores obtuvieron puntuaciones de *F1* en el conjunto de prueba (20 %) de un 94,89 % en segmentos de audio y del 92,02 % ea nivel de audio empleando CNN utilizando espectrogramas *log Mel* como características.

En [10], Zhou et al. abordan la detección de contenido pornográfico en archivos de audio empleando un modelo basado en CNN y la investigación de refinamientos como mecanismos de atención, métodos de *pooling* , *label smoothing*, *warmup* y *knowledge distillation*. El artículo presenta una evaluación exhaustiva en un conjunto de datos grande y recién recopilado (224.127 audios pornográficos y 274.206 audios normales), logrando una exactitud del 97,19 % con la combinación de los refinamientos nombrados anteriormente.

Finalmente, Liu et al[11] proponen un método novedoso para la detección de contenido pornográfico en audio, superando las limitaciones de los algoritmos existentes. Introducen una característica complementaria que combina *log Mel*, MFCC y

GFCC (Coeficientes Cepstrales de Frecuencia de Gammatone, del inglés *Gammatone Frequency Cepstrum Coefficient*. Su enfoque utiliza la arquitectura*Dual-Path Fused Transformer Net* (DPFTNet) y Random Forest para la clasificación. Recolectaron 7942 muestras de datos de vídeos de *streams* pornográficos de Internet. Estos vídeos fueron convertidos a audio y divididos en archivos de audio de 10 segundos. Finalmente, obtuvieron 13.338 archivos de audio de un solo canal. Las métricas de evaluación alcanzaron un 93,20 % de exactitud y un 93,56 % de puntuación F1, demostrando su eficacia.

En la clasificación de audio sensible, la mayoría de trabajos no han explorado métodos de aumento de datos sobre el propio audio (añadiendo ruido aditivo gaussiano, cambio del nivel del tono, etc) o sobre los espectrogramas de Mel o MFCC (enmascarando ciertas regiones de la frecuencia o del tiempo de los espectrogramas). Dicho aumento de datos, ayudaría a generalizar mejor. Además, sería la primera vez que se utiliza el conjuntos de *Pornography-2K*, ya que los trabajos que han utilizado este conjunto de datos, han empleado un subconjunto para el entrenamiento y para el test. Hasta ahora, ningún estudio que aplique Aprendizaje por Transferencia (AT) para detectar contenido sensible en audio a través de espectrogramas ha investigado el impacto del ajuste fino (del ingles *fine-tuning* de las capas del modelo en el rendimiento óptimo del conjunto de entrenamiento. En su mayoría, los trabajos se limitan a utilizar modelos preentrenados únicamente para extraer características, sin explorar la posibilidad de entrenar algunas capas y evaluar cómo afecta su rendimiento. El objetivo del uso de AT es doble: en primer lugar, reducir el tiempo de entrenamiento; y en segundo lugar, examinar el impacto del ajuste fino en el rendimiento, comparándolo con el estado del arte.

III. Detección de contenido sensible en audio y vídeo

En esta sección se describe paso a paso el enfoque de la propuesta de detección de contenido sensible utilizando espectrogramas y AT con CNN (modelo *ConvNeXt V2 femto* [12]). Para ello, en esta sección se presenta la metodología empleada, para luego describir el conjunto de datos y el preprocesamiento realizado y finalmente las métricas de evaluación.

III-A. Espectrogramas

Para digitalizar la información presente en los archivos de audio, se emplea comúnmente un proceso de muestreo que captura puntos de la señal de audio misma a intervalos regulares a lo largo del tiempo. Básicamente, se toman mediciones de la amplitud de la señal de audio en intervalos específicos y se convierten en valores digitales que pueden ser almacenados y procesados por dispositivos digitales. La velocidad de este proceso de muestreo puede variar, siendo comúnmente de 44,1 kHz (equivalente a 44,100 muestras por segundo) o 16 kHz (16,000 muestras por segundo). Una vez completado este proceso, se obtiene una representación digital de la señal de audio en forma de onda, la cual puede ser interpretada, modificada y analizada mediante software especializado (Fig. 1).

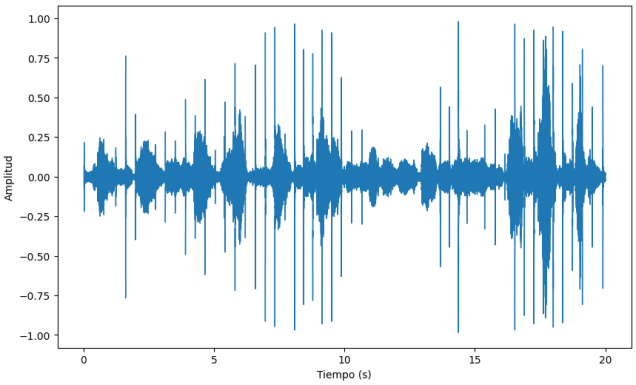

Figura 1. Representación digital de la señal de audio en forma de onda

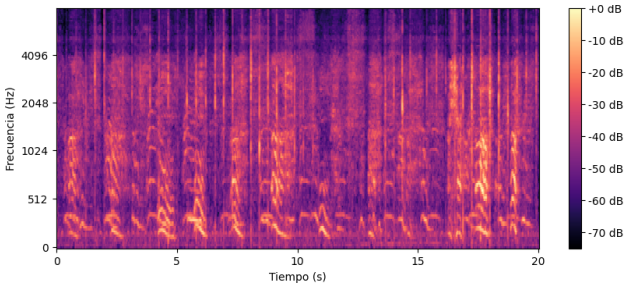

Figura 2. Espectrograma de *log Mel*

Gracias a la Transformada Rápida de Fourier (TRF, del inglés FFT), se puede descomponer la señal de audio en frecuencias individuales y amplitudes asociadas a cada frecuencia. En esencia, transforma la representación de la señal del dominio del tiempo al dominio de la frecuencia. El resultado de esta transformación se conoce como espectro. Dado que la mayoría de señales de audio varían con el tiempo (señales no periódicas), es necesario una representación del espectro que tenga en cuenta esta variación temporal. Por otro lado, la Transformada de Fourier de Tiempo Corto (TFCT, del inglés STFT en inglés) permite descomponer una señal de audio en pequeños segmentos superpuestos, permitiendo una representación temporal y frecuencial conocida como espectrograma (Fig.2. El espectrograma, es una herramienta muy útil para analizar y visualizar señales de audio en el dominio del tiempo y la frecuencia. Para su representación, el eje vertical se transforma a una escala logarítmica, mientras que la dimensión de color se convierte a decibelios, lo que equivale a una escala logarítmica de amplitud. Esta adaptación se realiza porque los humanos perciben una gama estrecha y concentrada de frecuencias y amplitudes, lo que se refleja en el espectrograma de Mel logarítmico (*log Mel*).

III-B. Metodología

Dado que los datos de audio presentan una gama diversa de sonidos, que van desde el habla humana hasta otros ruidos ambientales, resulta esencial emplear características acústicas capaces de distinguir entre datos de audio sensibles y no sensibles. Para este fin, se han utilizado las características MFCCs, que son ampliamente reconocidos en la detección del habla, como los espectrogramas logarítmicos de Mel. Estos últimos contienen una cantidad más extensa de información

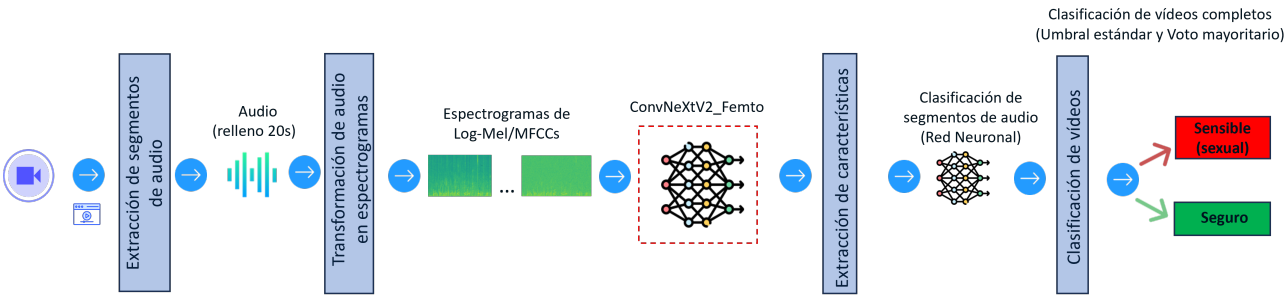

Figura 3. Flujo de trabajo para la clasificación de audios y vídeos.

acústica, lo que les permite retener una mayor cantidad de artefactos en frecuencias altas y bajas. Dichos artefactos o características retienen sonidos que no son típicos del habla, como gemidos o sonidos de movimientos corporales, pero resultan fundamentales en la detección de audio sensible. Tanto los MFCCs como los espectrogramas logarítmicos de Mel son ampliamente conocidos y han sido utilizados en la reconocimiento del habla (*speech recognition*), pero su aplicación, junto con el AT, en la detección de audio sensible aún es una novedad. Una vez que se ha presentado como se pueden representar una forma de onda (audio) como una imagen, se va a detallar la metodología seguida para la clasificación de vídeos utilizando únicamente características acústicas. El principal objetivo de este trabajo es clasificar segmentos de audio y vídeos completos sensibles mediante espectrogramas empleando uno de las ultimas arquitecturas de CNN del estado del arte (*ConvNeXt V2* previamente entrenada en un conjunto de datos que originalmente no está diseñado para clasificar espectrogramas, como el conjunto de datos *ImageNet*. Para ello, tras revisar los últimos trabajos del estado del arte, primero se extrajeron de cada vídeo segmentos de audio de 20 segundos con solapamiento de 1 segundo entre audios (para no perder información entre audios consecutivos). En una segunda etapa, se obtuvieron los espectrogramas de *log Mel* y MFCCs mediante diferentes parámetros (y por tanto diferentes espectrogramas), como el numero de bandas de mel (bins), el tamaño de la ventana de análisis en muestras de audio (Número de Puntos de la Transformada Rápida de Fourie, del inglés Number of Points of the Fast Fourier Transform (*nfft*) y la longitd de salto (*hop lenght*) Respecto a el modelo pre-entrenado utilizado, dentro de la familia de modelos de *ConvNeXt V2*, el modelo de *ConvNeXt V2 femto* es el segundo modelo con menos parámetros (5,2 millones), comparable a los 5,3 millones de parámetros del modelo empleado en [13] para la detección de contenido sensible sobre el mismo conjunto de datos (*Pornography-2K*. La arquitectura *ConvNeXt V2* incorpora una estructura de *autoencoder* enmascarado completamente convolucional y una capa de normalización de respuesta global (GRN), lo que mejora el rendimiento en múltiples pruebas comparativas. Para la etapa de entrenamiento, se utilizaron diferentes métodos de aumento de datos para mejorar la generalización del modelo. Entre ellos, se aplicaron diferentes aumentos en el audio (por ejemplo añadir ruido aditivo *gaussiano* o el cambio de tono) como en el espectrograma (enmascarar parte de la frecuencia, eje Y, como del tiempo, eje X). Finalmente, se utilizaron dos métodos para la clasificación a nivel de vídeo a partir de sus correspondientes predicciones a nivel de segmento de audio. El primero de ellos fue un umbral estándar, en el que la probabilidad de predicción a nivel de vídeo es igual a la media de sus respectivas predicciones a nivel de segmento. Una forma de onda de audio es sensible (clase 1) si la probabilidad es superior al umbral, o segura (clase 0) si es menor o igual a 0,5. El segundo método fue el voto mayoritario, en el que una clase se asigna a nivel de vídeo en función de la clase más frecuente entre las predicciones de los segmentos individuales. En este método, la clase asignada a un vídeo se determina por la clase que obtiene la mayoría de votos entre los segmentos de audio que lo componen.

En la *Fig.3* se puede observar las etapas descritas anteriormente con mas detalle.

III-C. Extracción de segmentos de audio

Para la extracción de segmentos de audio de vídeos del conjunto de datos *Pornography-2K*, se utilizó la librería *MoviePy* [14]. Para cada vídeo, se calculó el número de segmentos de audio necesarios basados en una duración máxima de 20 segundos por segmento. Posteriormente, se determinó su tiempo de inicio y finalización, se extrajeron los segmentos de audio y se guardaron finalmente como un archivo *WAV* con una tasa de muestro de 16Khz y con una precisión de 16 bits.

III-D. Conjunto de datos

Tras la extracción de segmentos de audio, se obtuvo un conjunto de datos formado por 7.795 y 18.498 segmentos de audio seguros y sensibles respectivamente (*Tabla I*). Cabe destacar que no se pudo extraer segmentos de audio de 24 vídeos, 12 vídeos seguros y 12 sensibles, ya que no estos vídeos no contenían pista de audio).

Tabla I
CONJUNTO DE DATOS NPDI *Pornography-2k*: NÚMERO SEGMENTOS DE AUDIO DE 20S EXTRAÍDOS POR CADA CLASE.

Clase	Nº segmentos de audio	Nº de vídeos
Segura	7.794	988
Sensible	18.498	988
Total	26.292	1.976

Como se puede observar, el conjunto de datos está desequilibrado, ya que la duración de los vídeos sensibles del conjunto de datos *Pornography-2K* es mayor que la de los vídeos seguros. Es crucial tener en consideración este desequilibrio al seleccionar la métrica adecuada, la cual se explicará más adelante.

III-E. Preprocesamiento

En primer lugar, se preparó el espectrograma de Mel/MFCC para su utilización en un modelo de AP. Posteriormente, se procedió a normalizar el espectrograma de Mel/MFCC para garantizar que los valores se encuentren dentro de un rango específico. Primero, se verifica si la diferencia entre el valor máximo y el valor mínimo del espectrograma es diferente de cero. Si es así, significa que hay variación en los valores y se procede a normalizar dividiendo cada valor del espectrograma por la diferencia entre el máximo y el mínimo, después de haber restado el mínimo a todos los valores (es decir, normalización *min-max*). Esto asegura que todos los valores estén en el rango de 0 a 1. Si la diferencia entre el máximo y el mínimo es igual a cero (lo que indica que todos los valores son iguales), se omite la división y simplemente se resta el mínimo a todos los valores. Esto evita una división por cero y mantiene los valores iguales, lo que podría suceder en casos de espectrogramas con valores constantes. Luego, se ajusta el rango de valores para que estén en el rango [0, 255], que es el rango típico para valores de píxeles en imágenes. Esto se hace multiplicando por 255. Después, se replica el único canal del espectrograma de Mel/MFCC para crear una imagen RGB, que es necesaria para que el modelo pre-entrenado *ConvNeXt V2 femto* pueda procesarla. A continuación, se redimensiona la imagen para que tenga un tamaño específico (224x224 píxeles) utilizando el método de interpolación bicúbica. Después de la redimensión, se normaliza la imagen de acuerdo con la media y la desviación estándar típicas de las imágenes en el conjunto de datos de *ImageNet*, lo que ayuda al modelo a procesar la imagen de manera eficiente y precisa.

En las Fig.4 y Fig.5 se pueden observar las representaciones visuales de *log Mel* y MFCC respectivamente. En cuanto al aumento de datos, en la Fig.6 se muestra las transformaciones de enmascaramiento de la frecuencia y del tiempo junto con el ruido blanco y cambio de tono). Al aumentar los datos de entrenamiento con perturbaciones del tono, los modelos de detección de audio o de reconocimiento del habla se vuelven más robustos ante voces con diferentes rangos de tono. El enmascaramiento temporal consiste en seleccionar aleatoriamente un segmento continuo de la señal de audio y sustituirlo por silencio. Por otra parte, el enmascaramiento de frecuencias, por su parte, consiste en enmascarar un intervalo de frecuencias consecutivas en el espectrograma de audio. Esta técnica ayuda al modelo a ignorar componentes de frecuencia irrelevantes que pueden surgir del ruido o de interferencias de fondo.

III-F. Métricas de rendimiento

El rendimiento se ha expresado en términos *F1-measure* y de la exactitud (*accuracy*), priorizando *F1-measure macro* durante la búsqueda de los mejores hiper-parámetros durante la validación cruzada (5 *fold*).

La métrica *F1-macro* es comúnmente utilizada en la evaluación de modelos de clasificación, especialmente cuando se enfrentan conjuntos de datos desequilibrados, es decir, cuando algunas clases tienen muchos más ejemplos que otras. De esta manera, se calcula el *F1* para cada clase por separado y luego calcula el promedio no ponderado de estas medidas

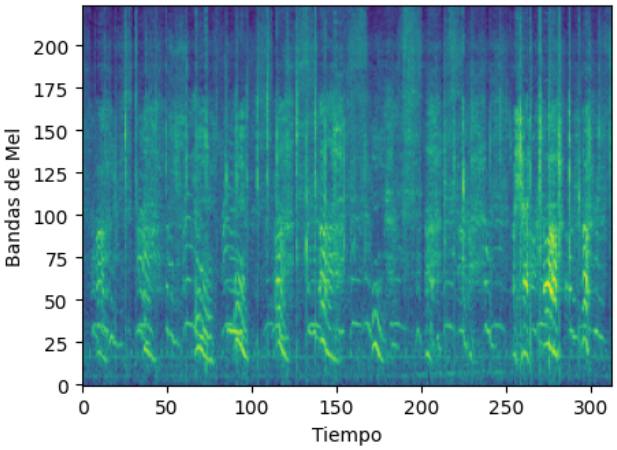

Figura 4. Espectrograma de *log Mel*.

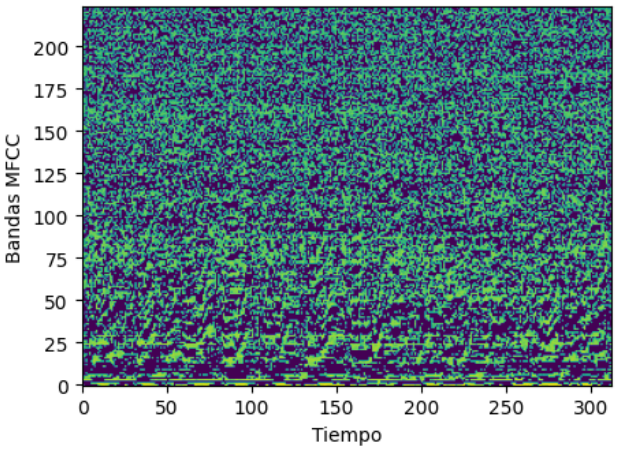

Figura 5. Espectrograma MFCC.

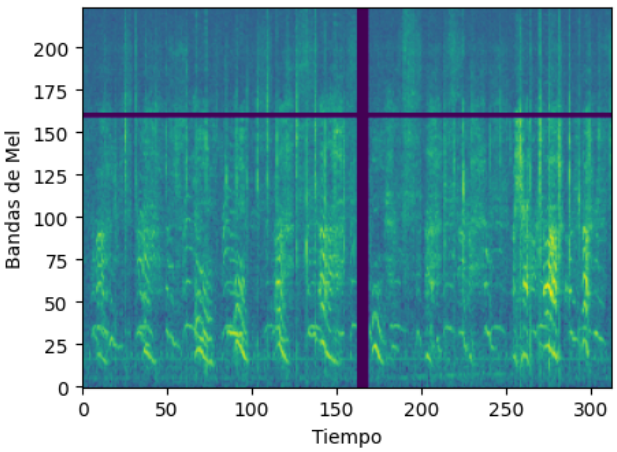

Figura 6. Aumento de datos sobre espectrograma de *log Mel*.

F1 individuales. Esto significa que cada clase contribuye de manera igual al resultado final, independientemente de su tamaño o distribución en el conjunto de datos.

Sean los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN), respectivamente, la exactitud (*accuracy*) de clasificación , *F1-measure*

y *F1-macro* se definen como sigue:

$$Exactitud = 100 \cdot \frac{VP + TN}{VP + FP + VN + FN} \quad (1)$$

$$Precisión = \frac{VP}{VP + FP} \quad (2)$$

$$Sensibilidad = \frac{VP}{VP + FN} \quad (3)$$

$$F_\beta = (1 + \beta^2) \cdot \frac{Precisión \cdot Sensibilidad}{\beta^2 \cdot Precisión + Sensibilidad} \quad (4)$$

donde β = 1 para *F1-measure*.

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^{N} F1_i \quad (5)$$

IV. Experimentos

El código para llevar a cabo todos los experimentos fue escrito en *Python 3.11*. El entrenamiento (ajuste fino de los modelos) se realizó con *PyTorch* 2.1 en un ordenador Lenovo con un procesador *Intel* de 13ª Generación a 2 GHz de frecuencia, *32 GB de RAM* y con una *GPU Nvidia RTX 3050* con 8 GB de memoria *VRAM*.

IV-A. Detalles de la implementación

En cuanto al aspecto técnico, primero se dividió de forma estratificada el conjunto de datos *Pornography-2K* en conjunto de entrenamiento (80 %) con 1.580 vídeos y conjunto de prueba (20 %) con 396 vídeos. Posteriormente, se obtuvieron los segmentos de audio específicos de cada conjunto de datos, dando lugar a 20.922 y 5.370 segmentos de audio de entrenamiento y prueba respectivamente. Como se ha comentado en el apartado anterior, el tamaño recomendado de entrada del modelo *ConvNeXt V2 femto* es de 224x224, por lo que las imágenes (espectrogramas) se re-dimensionaron a dicha resolución por cada canal uno de los canales (RGB). Para la construcción del modelo de referencia o *baseline* se utilizó el *ConvNeXt V2 femto* por defecto, congelando todas las capas salvo la parte superior del modelo, compuesto por una Red Neuronal *Feedforward*. Además, se utilizó el optimizador *AdamW* con un *learning rate* de 0.0005 y un decaimiento exponencial del aprendizaje del 15 % por época. Para el entrenamiento con *ConvNeXt V2 femto*, fue necesario descargarse el modelo de *timm* con los pesos de *ImageNet* sin incluir la capa final o *top* del modelo. Para el ajuste fino se realizaron experimentos ”descongelando”las últimas 15 y 30 capas. En cuanto a la arquitectura, se añadió una capa *GlobalAveragePooling2D* y una *Fully Connected Layer* (FCL) compuesta por 3 capas con 256, 128 y 64 neuronas respectivamente para la clasificación final. Para una mayor rapidez en la convergencia, se añadió una capa *BatchNormalization* antes de la función de activación *ReLu* y finalmente una capa de regularización *dropout* entre cada una de las capas para evitar el problema de sobreajuste.

Para mejorar la capacidad de generalización del modelo, se establecieron las siguientes transformaciones o aumento de datos (*data augmentation*): enmascaramiento de tiempo y frecuencia (con valores de 10 y 5 respectivamente), cambio de tono (entre -0,3 y 0,3), ruido aditivo *gaussiano* (entre 0,05 y 0,1). Los valores anteriores junto con el numero de bandas de Mel (224), el tamaño de la ventana de análisis en muestras de audio (2048/1024) número de longitud de salto (1024/512) se seleccionaron tras realizar varias pruebas con diferentes espectrogramas (valores altos o bajos suponían transformaciones bastante significativas en la imagen que impedían reconocer el propio contenido).

IV-B. Experimentos para la detección de contenido sensible

Respecto a los hiper-parámetros (HP) del entrenamiento (validación cruzada) en los experimentos, se establecieron tras realizar diversas pruebas unos HP base (Tabla II): optimizador AdamW con un planificador *ExponentialDecay* con un *learning rate* de 0.0005 y un *decay* de 0.95, con un tamaño de lote o *batch size* de 128. Además, se aplicó la ponderación por clases (*class weights*) para paliar el efecto de los datos desequilibrados. Además, se utilizaron diferentes métodos de regularización para evitar el sobreajuste durante el entrenamiento, como la parada temprana (*early stopping*), *drop out* y suavizado de etiqueta (*label smoothing)*.

Por otro lado, como se ha dicho anteriormente, también se probaron diferentes valores en los principales parámetros que controlan la extracción de *log Mel* y MFCC, como se muestra (Tabla III).

Tabla II
HIPER-PARÁMETROS UTILIZADOS EN EL ENTRENAMIENTO CON EL MODELO *ConvNeXt V2 femto*

Hiperparámetro	Modelo base (*Baseline*)	Ajuste-fino (*Fine-tuning*)
Optimizador	AdamW	AdamW
Decaimiento de pesos	0.01	0.01
Tasa de aprendizaje	decaimiento exponencial (lr=0,0005, decay=0,95)	decaimiento exponencial (lr=0,0005, decay=0,95)
Tamaño de lote	128	128
Drop out	0,5	0,5
Parada temprana	Si	Si
Suavizado de etiqueta	0.1	0.1
FCL (3 capas)	[256,128,64]	[256,128,64]
Aumento de datos	No	Si
Capas entrenables	No	Si
Ponderación por clases	Si	Si

Tabla III
DISTINTOS VALORES EMPLEADOS PARA LAS BANDAS DE MEL Y LA LONGITUD DE SALTO DURANTE EL ENTRENAMIENTO CON EL MODELO *ConvNeXt V2 femto*

Hiperparámetro	Nº bandas de Mel y MFCC	NFFT	*Hop lenght*
Combinación 1	224	2048	1024
Combinación 2	224	1024	512

Finalmente, en la Tabla V, se presentan las distintas configuraciones de los experimentos llevados a cabo y los resultados obtenidos mediante validación cruzada. Para una visualización más gráfica y clara de la variabilidad en los resultados (*F1-measure*), se proporciona en la Figura 7 un diagrama de cajas.

V. Resultados y discusiones

En esta sección, se presentan los resultados finales obtenidos con el mejor modelo. A la vista de los resultados obtenidos en la Fig. 7, se puede observar como el modelo del experimento 4 (destacado con un color mas oscuro) es el

Tabla IV
CONFIGURACIÓN DE LOS EXPERIMENTOS

Exp.	Espectrogramas	HP para obtener espectrogramas	Capas entrenables	Aumento de datos	F1(%)
1	*log Mel*	Combinación 1	0	No	81,72
2	*log Mel*	Combinación 1	15	No	84,80
3	*log Mel*	Combinación 1	30	No	85,62
4	*log Mel*	Combinación 1	30	Si	85,05
5	*log Mel*	Combinación 2	30	Si	85,39
6	MFCC	Combinación 1	30	Si	80,62
7	MFCC	Combinación 2	30	Si	80,72

mas robusto y con un *F1-measure* muy parecido al modelo con *log Mel* y aumento de datos (experimento 5). Tras la elección del mejor modelo, en la Tabla *V* se muestran los resultados sobre el conjunto de datos de prueba. Para generar las predicciones finales (segura/sensible), se aplicó la función *argmax* sobre las probabilidades generadas por el modelo (dichas probabilidades se calcularon aplicando una función *softmax) sobre la salida del modelo*). Esto implica que la clase con la probabilidad más alta en la salida del modelo se selecciona como la predicción final. Como se puede observar, se lograron resultados superiores al 85 % en la puntuación de *F1-measure*, superando los resultados obtenidos en la validación cruzada mostrada en la Tabla IV. Esto sugiere que el modelo construido no sufre de sobreajuste (es decir, es robusto frente a datos no vistos) y su rendimiento está cercano a los obtenidos en otros estudios sobre conjuntos de datos similares.

El análisis de los resultados revela un mayor rendimiento utilizando 2048 NFFT y un *hop lenght* de 1024 en espectrogramas de *log Mel* frente a MFCCs, dando lugar a un modelo con un rendimiento destacable en la clasificación de instancias tanto positivas como negativas. Las métricas de precisión y sensibilidad muestran una capacidad consistente para predecir correctamente ambas clases. Sin embargo, es importante destacar que, a pesar de la alta precisión general del 76 % para la clase negativa (segura), existe cierta dificultad en clasificar correctamente algunas instancias como negativas. Esto sugiere una posible área de mejora en la capacidad del modelo para discernir con mayor precisión entre las instancias negativas y positivas (estudiando un *threshold* diferente o añadiendo mas muestras de segmentos de audio de la clase segura). A pesar de esta posible limitación, las puntuaciones *F1-measure* indican un equilibrio satisfactorio entre la precisión y la sensibilidad para ambas clases, lo que sugiere una capacidad general del modelo para clasificar con precisión. La matriz de confusión (Fig.8) proporciona una descripción detallada de cómo el modelo realiza las clasificaciones, destacando tanto las predicciones correctas como las incorrectas para cada clase. Se observa como el *F1-measure*, muestra un valor de 81,4 % para la clase 0 (audio seguro) y 91,5 % para la clase 1 (audio sensible), lo que indica un buen equilibrio entre precisión y sensibilidad en ambas clases.

En cuanto a la clasificación de vídeos completos, en las tablas *VI y VII* se muestran los resultados obtenidos mediante voto mayoritario (predicción que más se repite) y el método de la media de las probabilidades de cada segmento de audio por vídeo, con un umbral estándar de 0,5 (50 %).

Mediante el método de la Tabla VII se ha obtenido un rendimiento superior (90,4 %) de *F1-macro* respecto al obtenido

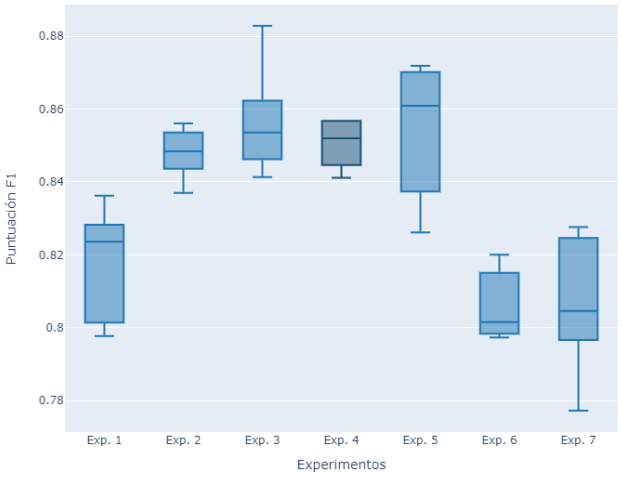

Figura 7. Gráfico de cajas de la validación cruzada.

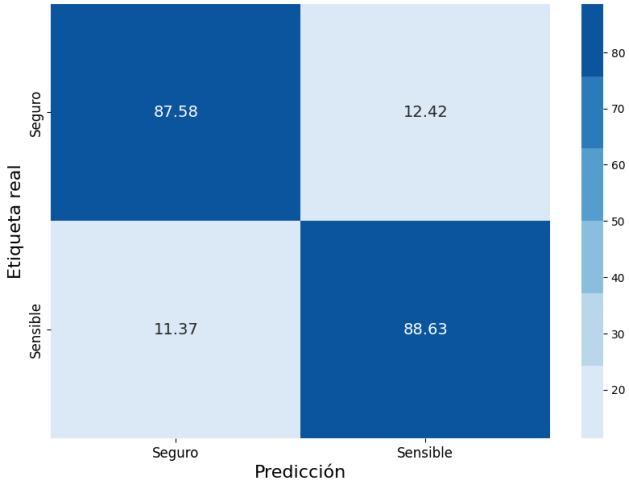

Figura 8. Matriz de confusión sobre el conjunto de prueba de segmentos de audio

mediante el voto mayoritario VI.

De las 396 muestras de datos del conjunto de pruebas de vídeos, solo 38 fueron clasificadas incorrectamente. Para investigar más a fondo las posibles causas de estas clasificaciones erróneas, se realizó un examen manual de algunos resultados de la prueba (segmentos de audio que fueron predichos incorrectamente). Se observó que los FN pueden atribuirse a la presencia de música u otros ruidos continuos de fondo en algunas partes de los datos de audio. En otros casos, los datos de audio carecen de sonidos relacionados con el contenido adulto y consisten principalmente en sonidos ambientales, como los de tormentas y agua. Estas características de audio llevan al modelo a clasificar erróneamente estos segmentos como seguros (no pornográficos).

Por otro lado, los FP suelen comprender sonidos repetitivos, como el tic-tac de un reloj o el sonido del agua, así como sonidos humanos repetitivos, como suspiros y pesadez después del ejercicio. Estos tipos de sonidos pueden hacer que las características derivadas del audio seguro se asemejen de manera engañosa al audio pornográfico, lo que resulta en

clasificaciones incorrectas por parte del modelo.

En relación al estado del arte presentado en la sección II, el único trabajo con el que se podría hacer una comparación justa (conjunto de datos parecido) es [8]. En dicho trabajo, obtuvieron sobre un conjunto de datos de prueba de 160 muestras, un 94,89 % de *F1-measure*. Esto resalta que, aunque en este trabajo se ha obtenido un rendimiento un 4 % menor, el conjunto de datos *Pornography-2K* incluye 1200 vídeos adicionales y más desafiantes de clasificar que el conjunto *Pornography-800* utilizado en [8].

Tabla V
RESULTADOS SOBRE SEGMENTOS DE AUDIO DEL CONJUNTO DE PRUEBA

Clase	Precisión (%)	Sensibilidad (%)	Puntuación F1 (%)
Segura	76	87,6	81,4
Sensible	94,6	88,6	91,5
Exactitud (*accuracy*)			88,3
F1 (media)	85,3	88,1	86,4

Tabla VI
RESULTADOS OBTENIDOS SOBRE VÍDEOS COMPLETOS DEL CONJUNTO DE PRUEBA MEDIANTE EL VOTO MAYORITARIO.

Clase	Precisión (%)	Sensibilidad (%)	Puntuación F1 (%)
Segura	89,2	91,4	90,3
Sensible	91,2	88,9	90
Exactitud (*accuracy*)			90,1
F1 (media)	90,2	90,2	90,1

Tabla VII
RESULTADOS OBTENIDOS SOBRE VÍDEOS COMPLETOS DEL CONJUNTO DE PRUEBA MEDIANTE EL UMBRAL ESTÁNDAR.

Clase	Precisión (%)	Sensibilidad (%)	Puntuación F1 (%)
Segura	89,6	91,4	90,5
Sensible	91,2	89,4	90,3
Exactitud (*accuracy*)			90,1
F1 (media)	90,4	90,24	90,4

VI. CONCLUSIONES

En este estudio se ha presentado una metodología para la construcción de un modelo para la detección de contenido sensible, específicamente pornografía adulta en ficheros de audio y vídeos, utilizando características acústicas. Para ello, se empleó una de las arquitecturas más destacadas en el campo de la Visión Artificial, conocida como *ConvNeXt V2*. Este trabajo constituye el primero en utilizar el conjunto de datos *Pornography-2k* en su totalidad para la detección de contenido sensible mediante características de audio, logrando una exactitud del 90,1 % y un 90,4 % de *F1-measure*. Además, mediante la aplicación de técnicas adecuadas de aumento de datos, se pudo reducir la variabilidad y aumentar la robustez del modelo, resultando en mejoras significativas en las métricas de rendimiento.

En cuanto al trabajo futuro, se observa que existen oportunidades para explorar y mejorar aún más el desempeño del modelo. Se sugiere investigar la aplicación del AT utilizando un modelo pre-entrenado en un conjunto de datos de espectrogramas, lo que podría proporcionar una mejora sustancial el rendimiento del modelo. Además, es crucial explorar otras estrategias de aumento de datos específicas para audio sensibles, dado que las técnicas actuales no son tan variadas como las disponibles para imágenes.

AGRADECIMIENTOS

Este trabajo se ha realizado con los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation), a través de la Cátedra "Ciberseguridad para la Innovación y la Protección Digital" INCIBE-UCM. Además, este trabajo ha contado con el apoyo del proyecto ALUNA (https://aluna-isf.eu/). Este proyecto ha recibido financiación del programa de investigación e innovación Horizonte 2020 de la Unión Europea en virtud del acuerdo de subvención nº 101084929. El contenido de este artículo no refleja la opinión oficial de la Unión Europea. Los autores son los únicos responsables de la información y las opiniones expresadas en él.

REFERENCIAS

[1] H. Adarsh and S. Sahoo, "Pornography and its impact on adolescent/teenage sexuality," *Journal of Psychosexual Health*, vol. 5, no. 1, pp. 35–39, 2023.

[2] C. S. Media, "Teens and pornography," Tech. Rep., 2022. [Online]. Available: https://www.commonsensemedia.org/research/teens-and-pornography

[3] L. Wang, J. Zhang, Q. Tian, C. Li, and L. Zhuo, "Porn streamer recognition in live video streaming via attention-gated multimodal deep features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 4876–4886, 12 2020.

[4] K. H. Song and Y.-S. Kim, "Pornographic video detection scheme using multimodal features," *Journal of Engineering and Applied Sciences*, vol. 13, no. 5, pp. 1174–1182, 2018.

[5] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Pornography classification: The hidden clues in video space–time," *Forensic science international*, vol. 268, pp. 46–61, 2016.

[6] K. Song and Y. S. Kim, "An enhanced multimodal stacking scheme for online pornographic content detection," *Applied Sciences (Switzerland)*, vol. 10, 4 2020.

[7] Z. Fu, J. Li, G. Chen, T. Yu, and T. Deng, "Pornnet: A unified deep architecture for pornographic video recognition," *Applied Sciences*, vol. 11, no. 7, p. 3066, 2021.

[8] H. Lovenia, D. P. Lestari, and R. Frieske, "What did i just hear? detecting pornographic sounds in adult videos using neural networks," in *Proceedings of the 17th international audio mostly conference*. Association for Computing Machinery, 9 2022, pp. 92–95.

[9] S. Avila, N. Thome, M. Cord, E. Valle, and A. De A. Araújo, "Pooling in Image Representation: The Visual Codeword Point of View," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[10] L. Zhou, K. Wei, Y. Li, Y. Hao, W. Yang, and H. Zhu, "Acoustic pornography recognition using convolutional neural networks and bag of refinements," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 840–845.

[11] S. Liu, R. Li, Q. Li, and J. Zhao, "Porn streamer audio recognition based on deep learning and random forest," *Applied Intelligence*, vol. 53, no. 15, p. 18857–18867, feb 2023. [Online]. Available: https://doi.org/10.1007/s10489-023-04491-x

[12] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.

[13] Povedano Álvarez, Daniel and Sandoval Orozco, Ana Lucila and García Villalba, Luis Javier, "Detección de contenido sexual explícito mediante técnicas de aprendizaje profundo," 2023, https://2023.jnic.es/wp-content/uploads/2023/06/PROGRAMA-JNIC-2023.pdf.

[14] Zulko, "Moviepy: Biblioteca de manipulación de videos y gifs en python," 2017, https://zulko.github.io/moviepy/.