COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

IMPLEMENTACIÓN DE LA INTELIGENCIA ARTFICIAL PARA LA GENERACIÓN DE VISUALES DINÁMICAS EN CONCIERTOS DE MÚSICA ELECTRÓNICA

Autor: Gabriel Llaneza Benosa

Director: Antonio Jesús Díaz-Cano Rincón

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

IMPLEMENTACIÓN DE LA INTELIGENCIA ARTIFICIAL PARA LA GENERACIÓN VISUALES DINÁMICAS EN CONCIERTOS DE MÚSICA ELECTRÓNICA

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2024/25 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Gabriel Llaneza Benosa Fecha: 07/06/2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Antonio Jesús Díaz-Cano Rincón Fecha: 07/06/2025

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

IMPLEMENTACIÓN DE LA INTELIGENCIA ARTFICIAL PARA LA GENERACIÓN DE VISUALES DINÁMICAS EN CONCIERTOS DE MÚSICA ELECTRÓNICA

Autor: Gabriel Llaneza Benosa

Director: Antonio Jesús Díaz-Cano Rincón

Madrid

Agradecimientos

Agradezco especialmente a mi director Antonio por su guía y apoyo durante todo el proyecto.

IMPLEMENTACIÓN DE LA INTELIGENCIA ARTIFICIAL PARA LA GENERACIÓN VISUALES DINÁMICAS EN CONCIERTOS DE MÚSICA ELECTRÓNICA

Autor: Gabriel Llaneza Benosa
Director: Antonio Jesús Díaz-Cano Rincón
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este proyecto desarrolla un sistema de software avanzado basado en inteligencia artificial para generar visuales interactivas que reaccionan en tiempo real a la música durante los conciertos. A través de modelos de IA y análisis de audio, el sistema ajusta las visuales al ritmo, las emociones y la estructura de la canción, logrando así automatizar su creación. Este enfoque facilita el acceso a artistas con menos recursos y ofrece experiencias visuales únicas en cada espectáculo en vivo.

Palabras clave: Inteligencia Artificial, Deep Learning, Visuales Reactivas, Análisis de Audio, TouchDesigner.

1. Introducción

En el ámbito de los conciertos de música electrónica, las visuales juegan un papel crucial en crear una experiencia envolvente para el público. Estas visuales, que suelen ser proyecciones artísticas en movimiento, están diseñadas para acompañar la música y aumentar la interacción sensorial durante el evento. Sin embargo, muchas de las técnicas actuales para su generación se basan en visuales pregrabadas, lo que limita su capacidad de reaccionar en tiempo real con la música o en visuales reactivas al audio tradicionales, que mejoran la variabilidad, pero no responden de manera adecuada o dinámica al contenido musical, lo que las hace poco flexibles y menos inmersivas.

Figura 1. Ombra Visuals Anima 1

Fuente: https://www.ombra.world/music

Este proyecto identifica una oportunidad para mejorar estas técnicas mediante la aplicación de inteligencia artificial (IA) consiguiendo generar visuales interactivas que responden de forma dinámica al ritmo, las emociones y los cambios de la música de manera inteligente.

2. Definición del proyecto

El proyecto se basa en la creación de un software sobre una arquitectura modular que combina análisis de audio, predicción mediante modelos de inteligencia artificial y generación visual en tiempo real. Utilizando fragmentos de audio como entrada, se extraen características acústicas relevantes que son procesadas por modelos de deep learning entrenados para identificar eventos musicales clave.

Los resultados generados por estos sistemas se integran en TouchDesigner, plataforma con la que se controla la generación de visuales en vivo, creando una experiencia inmersiva adaptada al contenido musical.

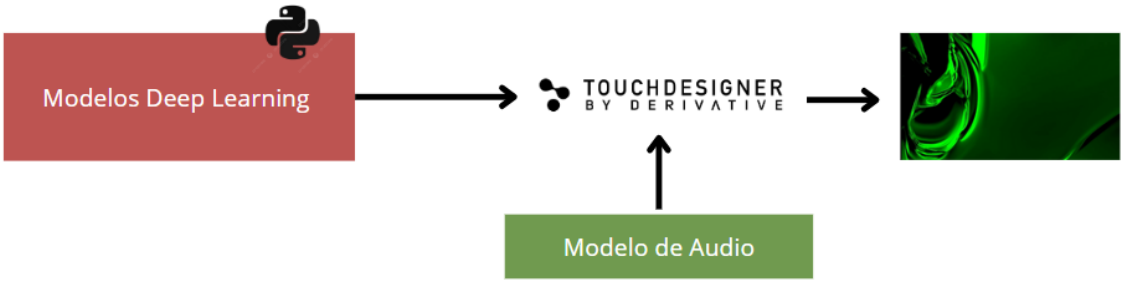

Figura 2. Esquema Simple del Proyecto

3. Descripción del modelo/sistema/herramienta

Se han desarrollado tres modelos de inteligencia artificial que conforman el núcleo del software, cada uno especializado en una tarea clave para lograr la generación de visuales reactivas e inteligentes. La combinación modular de estos modelos permite abordar distintos aspectos del audio y traducirlos en elementos visuales coherentes y dinámicos.

3.1. Modelo de predicción de género musical:

Este modelo de Deep Learning identifica el género musical de los fragmentos de audio, lo cual es clave ya que cada género tiene un estilo visual característico. Gracias a esta clasificación, las visuales se ajustan automáticamente para que coincidan con la música, mejorando la coherencia y la experiencia del espectador. Para ello, se emplea una arquitectura híbrida que combina una capa convolucional (CNN) para extraer características espaciales relevantes del espectrograma, seguida de una capa GRU (Gated Recurrent Unit)

que modela la información secuencial temporal del audio, capturando patrones que definen el género de manera precisa.

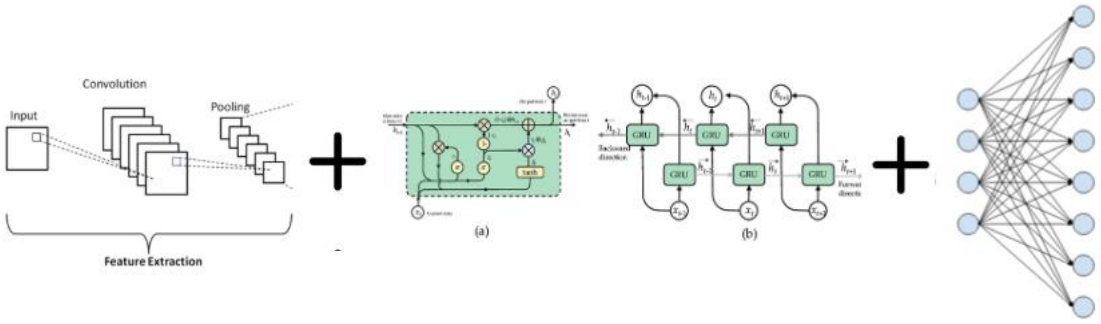

Figura 3. Arquitectura de red modelo de géneros

3.2 Modelo de clasificación de secciones musicales:

Con el fin de mejorar la estructura narrativa de las visuales, se ha desarrollado un modelo capaz de identificar las secciones o momentos clave dentro de una canción, tales como el build-up, el drop o el break. Reconocer estos segmentos permite adaptar la lógica visual para que refleje el desarrollo estructural de la música. Este modelo se basa en una arquitectura híbrida que combina una red ResNet-18 convolucional para la extracción de características visuales profundas, junto con una capa LSTM (Long Short-Term Memory) que modela las dependencias temporales y secuenciales dentro del audio.

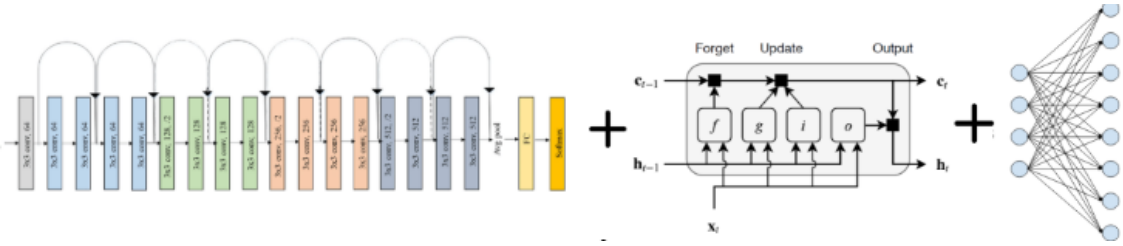

Figura 4. Arquitectura de red modelo de secciones

3.3 Modelo de reconocimiento de emociones musicales:

Finalmente, para dotar a las visuales de una dimensión expresiva más rica, se incluye un modelo que detecta las emociones predominantes en cada fragmento de audio, tales como energía, melancolía, alegría, entre otras. Estas emociones se traducen en parámetros visuales expresivos como la paleta de colores, el brillo o el contraste, contribuyendo a una experiencia inmersiva que conecta con el público a nivel sensorial y emocional. La arquitectura de este modelo combina una red convolucional para la extracción de características del audio con

un clasificador basado en un modelo de Random Forest, lo que permite mejorar la precisión y robustez en la detección emocional.

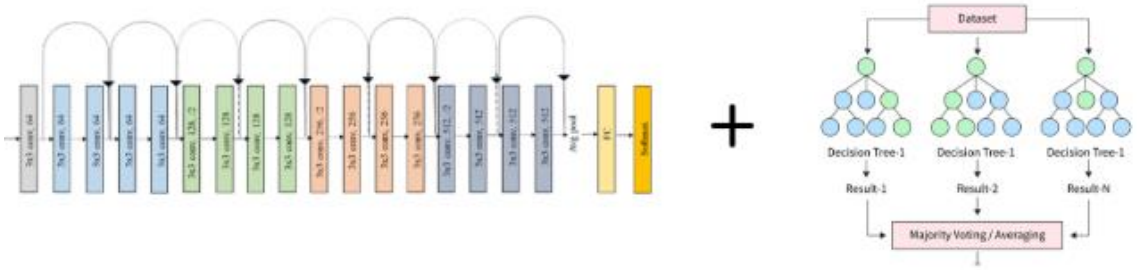

Figura 5. Arquitectura de red modelo de emociones

3.4 Sistema de análisis dinámico de audio:

Este sistema está diseñado para que las visuales puedan reaccionar de forma continua y fluida al ritmo, la energía y los cambios de intensidad de la música. Para lograr esta respuesta dinámica, el sistema procesa señales acústicas en tiempo real, extrayendo principalmente las de bombo, caja y ritmo, frecuencias bajas, altas y medias y densidad espectral y centroide que se traducen en movimientos y efectos visuales sincronizados.

En conjunto, estos cuatro modelos trabajan de manera integrada para analizar el audio desde diferentes perspectivas y traducir esta información en parámetros visuales que configuran la generación en tiempo real. Para la comunicación entre los modelos de inteligencia artificial y TouchDesigner, se utiliza el protocolo OSC (Open Sound Control), que permite transmitir de forma eficiente y en tiempo real los resultados y datos generados por los modelos hacia TouchDesigner.

Dentro del entorno de TouchDesigner, estos datos se reciben y se integran mediante nodos y scripts personalizados que interpretan cada señal proveniente de los modelos. Así, cada parámetro se mapea a atributos visuales específicos.

4. Resultados

Los modelos desarrollados, en general, han mostrado un rendimiento satisfactorio y una integración efectiva con el sistema de visuales reactivas. La combinación de CNN y GRU en la clasificación de géneros logró una accuracy superior al 64%, destacando especialmente la utilidad del enfoque probabilístico para generar visuales adaptativas y coherentes, aunque aún requiere cierta intervención manual en la preparación de plantillas.

El modelo de emociones, basado en Random Forest, capturó patrones emocionales relevantes en valencia y arousal, reflejando bien la evolución emocional de diferentes estilos musicales. Su integración con TouchDesigner permitió una visualización dinámica y armoniosa, aunque la escalabilidad y automatización del sistema podrían mejorarse.

Finalmente, el modelo de detección de pre-drop, a pesar de sus no tan buenos resultados en clasificación, fue efectivo para generar efectos visuales que amplifican la experiencia sonora en momentos clave. La principal mejora futura radica en ampliar y diversificar el conjunto de datos para aumentar su precisión.

Además, el sistema de análisis de audio responde de forma rápida y adecuada, logrando una buena sincronización con las visuales y mejorando la fluidez de la experiencia en vivo.

En resumen, los modelos cumplen con los objetivos funcionales y sientan una base sólida para futuros avances, demostrando viabilidad y potencial para espectáculos en vivo adaptativos y enriquecidos.

5. Conclusiones

Este Trabajo de Fin de Grado confirma la viabilidad de integrar inteligencia artificial en espectáculos visuales en tiempo real, pese a las limitaciones de tiempo y recursos. Se desarrolló un sistema funcional que utiliza modelos de Deep Learning para analizar audio y generar visuales reactivas durante actuaciones en vivo, cumpliendo la mayoría de los objetivos planteados.

El proyecto, innovador y multidisciplinar, sienta las bases para un software escalable y automatizado que facilite a artistas sin experiencia técnica el uso de visuales inteligentes. Las mejoras futuras incluyen ampliar los datos, automatizar el sistema y enriquecer la variedad estética, abriendo camino a una herramienta creativa que democratice las experiencias audiovisuales en directo.

6. Referencias

[1] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). "Convolutional recurrent neural networks for music classification".

[2] X. Jiang et al: Music Emotion Recognition Based on Deep Learning (2024).

[3] Zhao, R., & Zhang, Y. (2022). "Hybrid Bi-LSTM and CNN Model for Emotion Recognition in Music"

[4] Lee et al. "Automatic Music Genre Classification Using Convolutional and Recurrent Neural Networks.

[5] A. Vaswani “Attention Is All You Need” (2017)

[6] "Deep Embeddings and Section Fusion Improve Music Segmentation" Nieto y Bello.

[7] "Pop Music Highlighter: Marking the Emotion Keypoints” (Huang et al., con referencia a Bittner et al.)

[8] Sun, L., Zhang, Y., & Zhang, Q. (2020). CNN-LSTM based emotion recognition for music structure analysis. IEEE Transactions on Affective Computing.

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

La convergencia entre la inteligencia artificial y el análisis musical ha supuesto la creación de un nuevo paradigma experimental que ha cambiado drásticamente la forma en que entendemos, clasificamos, experimentamos o incluso creamos música. Utilizando técnicas avanzadas como redes neuronales convolucionales (CNN) o redes neuronales recurrentes (RNN) se ha podido demostrar el enorme potencial de esta tecnología para identificar patrones en audio, clasificar canciones o generar composiciones musicales.

El uso de estas técnicas significa una gran innovación para la industria musical ya que, por primera vez, las canciones dejan de ser solamente escuchadas y pasan a ser vistas. Esto implica que, a través del análisis mediante inteligencia artificial, las canciones se convierten en objetos más comprensibles, sus patrones se hacen identificables, su estructura se vuelve predecible y las emociones que nos despiertan pueden ser clasificadas y visualizadas.

Dentro del contexto de la música electrónica esta capacidad de análisis musical tiene un gran impacto pues este género siempre ha estado a la vanguardia en cuanto a la incorporación de tecnología y arte. La música electrónica en directo se caracteriza por crear experiencias inmersivas a través de sonidos, atmósferas y visuales. Las visuales, que juegan un papel clave en amplificar esta inmersión emocional, son proyecciones gráficas diseñadas para complementar la experiencia de la música en vivo. Normalmente se proyectan en pantallas y abarcan variedad de estilos como videos sincronizados, formas geométricas o animaciones de objetos en 3D.

Figura 6. Visuales DJ SnakeParis 1

Fuente: https://www.ombra.world/music

Sin embargo, la relación entre el audio y las visuales en muchos espectáculos sigue siendo predefinida y estática, con proyecciones pregrabadas o que simplemente acompañan a la música sin una adaptación en tiempo real. Además, algunos espectáculos que sí buscan una atmósfera visual en tiempo real necesitan de profesionales encargados de controlar visuales durante todo el show, con sus limitaciones y costes, o el uso de técnicas de análisis de audio y sincronización tradicionales que no ofrecen una interpretación profunda de la música y sus patrones.

Este proyecto busca ir un paso más allá al implementar estas técnicas de inteligencia artificial para la generación de visuales en tiempo real, brindando una experiencia más envolvente y personalizada para el público que el análisis tradicional de audio. A través de tres modelos basados en Deep Learning para la predicción de género musical, emociones y secciones de la canción crearemos un software que, integrado en la plataforma de generación de visuales dinámicas, Touchdesigner, buscará conseguir el objetivo explicado. De esta forma frente al modelo tradicional de análisis de audio, nuestro software será capaz de:

- Predecir el género de la canción para establecer unas visuales acordes a él (estilo, textura, velocidad de movimientos…)
- Predecir el flujo de emociones durante la canción para variar los colores, el brillo o la intensidad
- Predecir en qué momento nos encontramos de la canción para coordinar posibles cambios de visuales o de reacción de movimientos

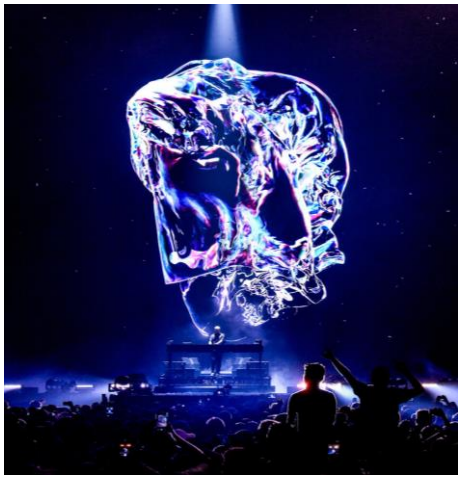

Figura 7. Ombra Visuals DJ Snake Paris 2

Fuente: https://www.ombra.world/music

Lo primero que detecta el funcionamiento del programa es el género de la canción de manera que, si el artista carga una canción de género Techno, automáticamente el software selecciona una plantilla de visuales acorde al género de la canción predicha. Para cada canción, aunque tengan el mismo género, se obtendrían visuales distintas pero similares en

cuanto a estilo. En el caso del Techno, las visuales serán minimalistas, patrones repetitivos, formas geométricas y líneas.

A medida que el artista reproduce la canción el modelo de emociones y el de detección de secciones van prediciendo, de forma continua, en ventanas de tiempo y en función de las emociones que se detecten en cada momento los colores, el brillo, ciertos efectos, o la intensidad, que pueden ir variando. Por ejemplo, si se detecta felicidad los colores pasarán a ser más brillantes.

El tercer modelo detecta en qué momento de la canción se ha entrado, por ejemplo, en el build-up de la canción (sección de la canción que precede al clímax y en la que la energía tiende a crecer de forma continua); en dicho punto las visuales se empezarían a mover de forma más rápida o bien a vibrar en esa preparación para el clímax.

1.1 MOTIVACIÓN DEL PROYECTO

El sector de las visuales en conciertos de música electrónica ha experimentado un notable crecimiento en los últimos años, reflejando una tendencia más amplia en la industria del entretenimiento en vivo y haciendo consolidarse a la electrónica como el género más innovador en este ámbito. Los conciertos han transcendido de simplemente lo sonoro para convertirse en experiencias visuales inmersivas. Artistas como Anyma o Eric Prydz invierten millones de euros en crear atmósferas visuales únicas, teniendo algunos de los espectáculos con más tickets vendidos mundialmente.

Este crecimiento en la industria ha impulsado la demanda de tecnologías innovadoras que mejoren la experiencia del público. Las visuales dinámicas y los efectos de iluminación avanzados son elementos clave para crear estas atmósferas únicas y memorables.

A pesar de la rápida expansión de este sector y su atractivo, todavía existe un amplio margen de desarrollo, especialmente con la reciente aparición de tecnologías como el Deep Learning y la inteligencia artificial.

Figura 8. Anima Event Martin Garrix Visuals

Fuente: https://www.ombra.world/music

La motivación principal de este proyecto es la mejora de dos áreas claves en el sector: Accesibilidad y Variabilidad; se apoya en el potencial de la inteligencia artificial para abordar dichas variables, al objeto de transformar las visuales en espectáculos en vivo.

Accesibilidad

A pesar de los avances, las visuales profesionales y reactivas siguen siendo, en su mayoría, inaccesibles para los artistas promedio. En un show de música electrónica de tamaño pequeño/medio, se requieren entre dos y tres personas para manejar las visuales en directo. En conciertos más grandes, el número de técnicos especializados puede ascender hasta 6-7 personas trabajando en tiempo real. Además, se necesita un equipo técnico y creativo aparte para la creación y la integración de las visuales con el hardware de los equipos durante cada evento.

Un ejemplo claro de esta falta de accesibilidad se encuentra en muchas discotecas medianas y pequeñas. Aunque la mayoría de estas están equipadas con pantallas, no se aprovechan adecuadamente. Los artistas pequeños que se presentan en estos lugares no cuentan con un sistema de visuales que complemente su show, ni con un equipo que les ayude a manejarlas

en tiempo real. Como resultado, las pantallas simplemente muestran el nombre del artista perdiendo una gran oportunidad de enriquecer la experiencia del público y de generar una atmósfera única. En conclusión, las visuales reactivas y profesionales continúan siendo inaccesibles para muchos artistas por falta de recursos técnicos.

Variabilidad:

Otro desafío importante es la falta de variabilidad en las visuales. La mayoría de los shows profesionales emplean sistemas de visuales pregrabadas y mínimamente manipuladas en directo, lo que genera que los conciertos del mismo artista se vuelvan prácticamente idénticos independientemente de la fecha o el lugar. Las visuales no responden de manera vibrante ni dinámica al ambiente del show, lo que reduce la sensación de autenticidad y la conexión con el público.

Este proyecto busca abordar esta limitación, aplicando variabilidad a las visuales, de modo que puedan responder de forma fluida y única a la energía del evento, las emociones y el entorno.

Es en este contexto donde nace el proyecto, para resolver estas necesidades y hacer que las visuales reactivas sean accesibles para artistas pequeños y medianos, que no cuentan con los recursos necesarios para contratar equipos especializados. A través de la implementación de inteligencia artificial y sistemas automatizados, esta solución, permite reducir la intervención humana al mínimo. Esto hace que la creación y ejecución de visuales en vivo sea más asequible, flexible y escalable, adaptándose tanto a shows pequeños como a grandes eventos.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

En el presente trabajo se abordan cuatro diferentes tecnologías específicas que se deberían introducir:

- **TouchDesigner:**

  TouchDesigner es un entorno de programación visual desarrollado por Derivative especializado en la generación de contenido interactivo, visuales y experiencias multimedia. Se basa en un sistema de nodos que permite manipular datos, imágenes y sonidos sin necesidad de escribir código, aunque también admite scripting de Python para personalización avanzada.

  TouchDesigner es ampliamente utilizado en la industria del espectáculo y de eventos en vivo ya que es una herramienta clave para la generación de visuales. Su integración con Python y su personalización permitirá integrar los modelos de inteligencia artificial. Su utilización generará las visuales.

- **Redes Neuronales Convolucionales (CNN):**

  Las redes convolucionales son un tipo de redes profundas especializadas en el procesamiento de datos con estructura espacial, imágenes, audio y video. Las CNN utilizan operaciones de convolución para extraer características de los datos mejorando el rendimiento en tareas de visión artificial y análisis de señales.

  Una red CNN está compuesta por varias capas:

  - Capas Convolucionales: Son el núcleo de las CNN y se encargan de aplicar filtros sobre los datos de entrada. Dichos filtros constituidos por matrices de pesos, detectan patrones en los datos, como texturas o bordes en las primeras capas y estructuras más complejas en las más profundas.
  - Función de Activación: ReLu, se utiliza para evitar linealidad en la red.
  - Pooling: Son capas que se encargan de reducir la cantidad de datos a procesar sin pérdida de información relevante.

- Capas Fully Connected o Densas: procesan las características extraídas para realizar la clasificación.

- **Redes Neuronales Recurrentes (RNN) y sus variantes (LSTM) y (GRU):**

  Las Redes Neuronales Recurrentes (RNN) son modelos diseñados para procesar datos secuenciales, como texto o audio, ya que pueden mantener memoria de pasos anteriores. Sin embargo, tienen dificultades para aprender dependencias a largo plazo. Para resolver esto, surgieron variantes como las LSTM (Long Short-Term Memory) y las GRU (Gated Recurrent Unit), que incorporan mecanismos de puertas para controlar el flujo de información, permitiendo el aprendizaje de secuencias más largas de forma estable y eficiente.

- **Espectrograma de Mel:**

  El Espectrograma de Mel es una representación visual del contenido de frecuencia de una señal de audio a lo largo del tiempo, basada en la escala Mel, que imita la percepción auditiva humana. Es ampliamente utilizado en tareas de procesamiento de audio, como reconocimiento de voz o análisis musical ya que utiliza una escala de frecuencias logarítmica en lugar de la lineal como un espectrograma de audio tradicional. Esto significa que, en lugar de dividir la frecuencia en segmentos iguales, utiliza una escala imitadora de la percepción del oído humano,

Capítulo 3. ESTADO DE LA CUESTIÓN

Este proyecto requiere un análisis exhaustivo del estado actual de la investigación en las dos tecnologías de las que hace uso, tanto en la de las visuales en vivo como en la del uso de aprendizaje profundo sobre audio.

3.1 GENERACIÓN DE VISUALES EN CONCIERTOS

Orígenes (1960s-1990s)

Las primeras configuraciones sincronizadas de visuales en conciertos eran muy simples y se diseñaban previamente para acompañar a la música, ejecutándose con control manual durante los conciertos logrando experiencias inmersivas pero muy limitadas. Bandas como Pink Floyd o The Greatful Dead fueron los primeros en experimentar con estas tecnologías, en su caso con proyectores analógicos.

Figura 9. Liquid Visuals The Greatful Death

Fuente: https://www.ombra.world/music

Primeros sistemas software (1990s-2000s)

En los años 90, con la mejora de la capacidad computacional se desarrollan sistemas software como Vjing que permite mezclar imágenes y videos en vivo. A partir de los 2000s, la aparición de softwares de creación de visuales dinámicas como Resolume o Touchdesigner

logró grandes avances en este campo generando visuales más sofisticadas y elaboradas. Además, la posterior incorporación de softwares de análisis de audio hizo posible que estas visuales respondiesen a la música de forma básica a través de audio-reactividad donde las visuales cambian con el volumen, los beats y la frecuencia principalmente.

Actualidad

Hoy en día la tecnología ha avanzado mucho, pero la base de las visuales sigue siendo la misma que la surge a partir de los 2000s. Hasta la fecha la mayoría de los conciertos y shows en vivo utilizan los siguientes tipos de visuales:

- Visuales Pregrabadas: utilizadas en grandes festivales y sin manipulación real.
- Visuales Reactivas: responden al audio con técnicas básicas como FFT pero limitadas en cuanto a mejora de la experiencia del espectador.
- Visuales manipuladas: Creadas antes del espectáculo, pero manipuladas en directo por técnicos.

Los principales softwares para su creación son Touchdesigner, Resolume y Notch que requieren operadores humanos.

Sin embargo, recientemente, la mejora de la inteligencia artificial está comenzando a transformar la creación de visuales, aunque de forma experimental, ya que la mayoría de sistemas basados en IA aún están en desarrollo y no se aplican en conciertos en vivo. Herramientas como Stable Diffusion empiezan a experimentar con la generación de visuales mediante inteligencia artificial, aunque no en tiempo real.

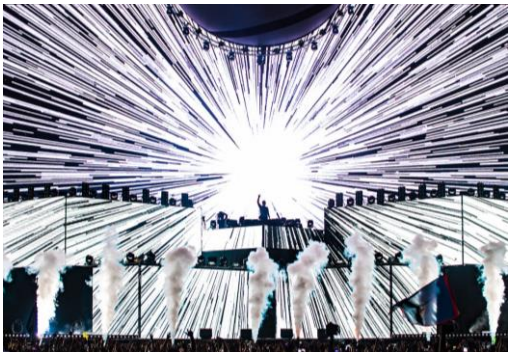

Figura 10. Visual Martin Garrix 2

Fuente: https://www.ombra.world/music

Durante este trabajo no se han encontrado referencias a sistemas de IA para visuales en tiempo real parecidos al de este proyecto, pero es probable que aparezcan en los próximos años. La tecnología enfrenta desafíos como la latencia y la capacidad computacional para sincronizarse con la música, lo que hace que este trabajo tenga en parte un enfoque experimental dado que aún no es lo suficientemente madura para un uso real.

3.2 TÉCNICAS DE APRENDIZAJE PROFUNDO EN AUDIO

En cuanto al estado actual del Deep Learning aplicado en música analizaremos los tres campos que utilizaremos en los tres modelos para la generación de visuales:

3.2.1 Clasificación de emociones en la música

El reconocimiento de emociones en la música ha sido un campo ampliamente estudiado en disciplinas como la psicología, la neurociencia y la teoría musical. La investigación en este ámbito se centra en cómo diferentes características del sonido, como patrones rítmicos, tonalidades y frecuencias, pueden evocar respuestas emocionales en los oyentes. Gracias al crecimiento de la inteligencia artificial, este campo ha avanzado considerablemente, con numerosos proyectos que emplean redes neuronales para analizar y clasificar emociones musicales.

Para las primeras investigaciones sobre clasificación de emociones en la música se utilizaban modelos de Machine Learning, como SVM, KNN o Random Forest. Estos modelos tenían limitaciones, principalmente la fuerte dependencia en la correcta selección de características extraídas de la señal de audio.

Sin embargo, con el rápido desarrollo de la inteligencia artificial, los modelos basados en Deep Learning han conseguido una mejora significativa en la clasificación de emociones musicales permitiendo el aprendizaje jerárquico de representaciones de audio.

El estudio de Choi et al [1] es clave en este ámbito introduciendo por primera vez una arquitectura híbrida de redes convolucionales y recurrentes (CRNN). Las Redes Neuronales Convolucionales (CNNs) se emplean para extraer patrones espectrales de los espectrogramas de audio, mientras que las Redes Neuronales Recurrentes (RNNs**)** capturan la dinámica temporal de las señales musicales.

Otro estudio relevante es el de X. Jiang et al [2], donde se compararon distintos tipos de arquitecturas profundas para clasificación de emociones. Los resultados indicaron que las redes híbridas CNN-LSTM eran especialmente efectivas para este tipo de tareas. O propuestas como la de Zhao, R., & Zhang, Y [3] que mostraban que la incorporación de un enfoque bidireccional en la LSTM mejoraba los resultados de predicción de emociones.

La idea es explorar estos modelos y crear uno ajustado para que se adapte a la música electrónica, priorizando el tiempo de procesamiento para que pueda funcionar en tiempo real. Aquí un ejemplo de la arquitectura de la red propuesta por Zhang:

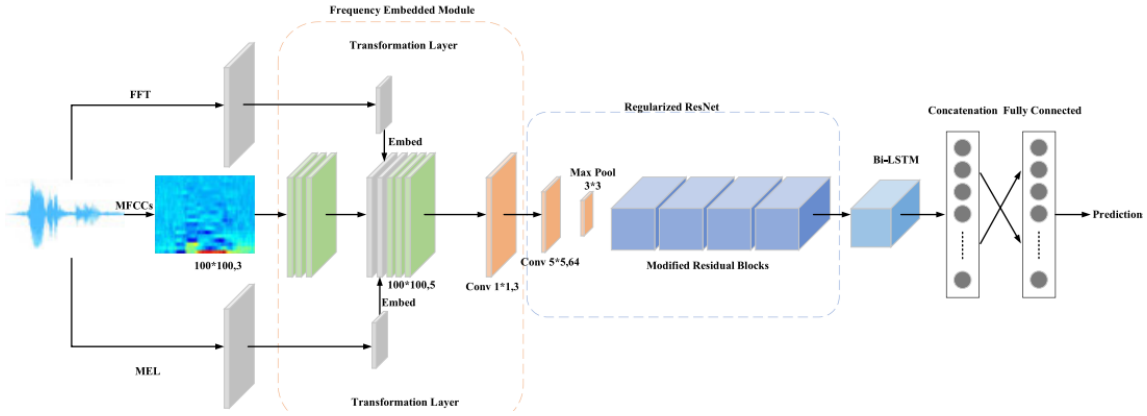

Figura 11: CNN-Bi-LSTM by Zhang [3]

Fuente:https://www.researchgate.net/publication/383237611_MUSIC_EMOTION_RECOGNITION_AND_CLASSIFICATION_USING_HYBRID_CNN-LSTM_DEEP_NEURAL_NETWORK

En cuanto al método de clasificación por emociones muchos de estos estudios utilizan una clasificación por valencia y arousal, un modelo de clasificación bidireccional propuesto por Russel (1980) que permite representar prácticamente cualquier emoción.

- Valencia: Indica si una emoción es positiva o negativa. Se mide en una escala de 0 a 1, donde el 0 representa emociones negativas como tristeza, ira o miedo y el 1 representa emociones positivas como alegría, felicidad o calma.
- Arousal: Mide la intensidad o activación de una emoción, también en una escala de 0 a 1, donde 0 representa emociones con baja activación, como relajación, calma o aburrimiento 1 representa emociones con alta activación, como excitación, tensión o euforia.

La combinación de ambos valores permite representar un amplio rango de emociones. Por ejemplo:

- Alta valencia y alto arousal → Alegría, entusiasmo.
- Alta valencia y bajo arousal → Relajación, satisfacción.
- Baja valencia y alto arousal → Ira, ansiedad.

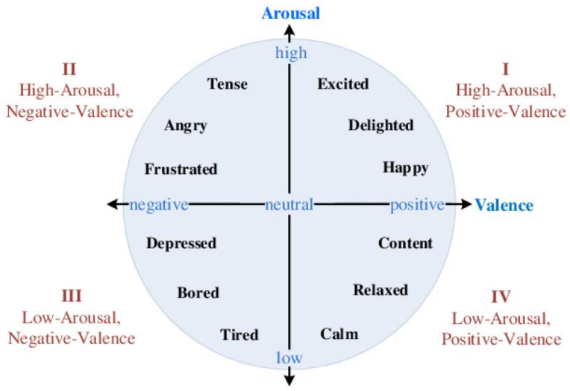

Figura 12: Modelo clásico Valencia y Arousal

Fuente: https://www.researchgate.net/figure/Two-dimensional-valence-arousal-space_fig1_304124018

3.2.2 Clasificación de género en la música

La clasificación de géneros musicales ha sido el área de la música más activamente investigada mediante el aprendizaje automático. Al igual que para la clasificación de emociones, los modelos más avanzados y que han surgido recientemente utilizan un modelo híbrido entre CNN y LSTM.

El modelo propuesto por Choi et al. [1] utiliza una red CRNN, donde las CNN extraen características del espectrograma y las LSTM procesan las dependencias temporales para realizar la clasificación. Otros estudios, como el de Lee et al. (2019) [4], han mejorado este enfoque con técnicas avanzadas de preprocesamiento y atención. Existen modelos muy efectivos con secuencias largas como el de Vaswani et al (2017) [5] que utiliza Transformers, pero requieren de mucha capacidad computacional haciéndolos menos prácticos y válidos para nuestro proyecto. Por lo tanto, la combinación de CNN y LSTM sigue siendo una opción eficiente para la clasificación de géneros musicales, proporcionando un buen equilibrio entre extracción de características y modelización temporal sin una alta demanda de recursos. Exploraremos estos modelos y los adaptaremos a nuestras necesidades.

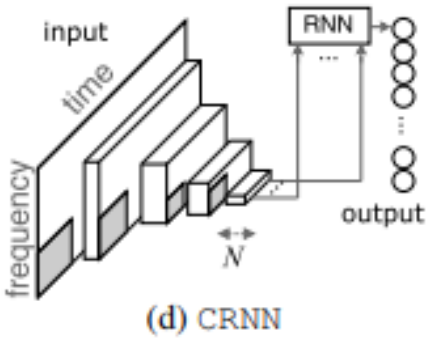

Figura 13: Arquitectura CRNN propuesta por Choi K [8]

Fuente:.https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/30083/Choi%20Convolutional%20recurrent%20neural%202017%20Accepted.pdf?sequence=1

3.2.3 Clasificación de secciones de la canción

La determinación de las secciones de la canción es un aspecto clave para el objetivo de nuestro proyecto. Recientemente, varias investigaciones han mejorado la capacidad de segmentar y clasificar las estructuras musicales utilizando redes neuronales, principalmente CNN y LSTM.

Nieto y Bello (2018) [6] propusieron el uso de redes CNN para detectar estructuras en canciones populares, logrando segmentaciones adecuadas en géneros tradicionales. Posteriormente, Bittner et al. (2019) [7] mejoraron este enfoque al integrar mecanismos de atención, lo que permitió una mayor precisión en la detección de transiciones musicales

complejas, un desafío recurrente en géneros con cambios abruptos o poco evidentes. Sun, L., Zhang, Y., & Zhang, Q. (2020) aplicaron un enfoque CNN-LSTM para la segmentación automática de canciones, mostrando la efectividad de combinar la extracción de características espaciales con la capacidad de modelar dependencias temporales. Este enfoque ha demostrado ser especialmente útil para identificar secciones como el verso, el coro y el puente en géneros más estructurados.

Sin embargo, la segmentación de canciones en géneros más complejos, como la música electrónica, sigue siendo un desafío debido a la naturaleza fluida y menos estructurada de las composiciones. Las transiciones en este género no son tan claras como en otros lo que supone un desafío importante para el proyecto.

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1. JUSTIFICACIÓN

Como se ha explicado previamente el sector de las visuales en conciertos de música electrónica ha evolucionado rápidamente, convirtiéndose en un componente esencial para la experiencia del público. Su crecimiento es exponencial y la inversión en esta tecnología no para de aumentar. Sin embargo**,** hoy por hoy no existe una solución accesible y profesional que permita a artistas pequeños y medianos disponer de visuales reactivas de calidad sin la necesidad de equipos técnicos especializados**.**

A pesar de los avances en inteligencia artificial y en generación de contenido audiovisual automatizado, la mayoría de los espectáculos siguen dependiendo de VJs, los técnicos encargados de la manipulación en tiempo real, lo que limita el acceso a las visuales profesionales a artistas con pequeño presupuesto. Además, la falta de variabilidad en los espectáculos es un problema recurrente: muchas visuales son pregrabadas y apenas se modifican entre conciertos, lo que reduce la originalidad y la conexión con el público.

Este proyecto soluciona ambas problemáticas a través de la inteligencia artificial**.** Al desarrollar un sistema basado en Deep Learning y modelos de análisis de audio en tiempo real**,** conseguimos que las visuales se generen y adapten automáticamente a la música y el ambiente del show sin intervención manual.

4.1.1. Diferenciación y ventajas competitivas

El desarrollo de este proyecto permite obtener una serie de mejoras en el mundo de los espectáculos musicales entre los que se destacan los siguientes puntos:

- **Reducción de costos y democratización del acceso**: Los artistas pequeños y medianos podrán contar con visuales de alta calidad sin la necesidad de contratar equipos técnicos costosos.
- **Generación automática y en tiempo real**: A diferencia de las soluciones tradicionales, donde las visuales deben ser preconfiguradas y manipuladas por un VJ, nuestro sistema permitirá respuestas instantáneas y dinámicas, basadas en el análisis en vivo de la música.
- **Mayor personalización**: Se evita la repetición de visuales en diferentes conciertos de un mismo artista, generando una experiencia única en cada show.
- **Integración eficiente con hardware existente**: Al utilizar protocolos estándar como OSC**,** el sistema podrá comunicarse fácilmente con plataformas como TouchDesigner, facilitando su adopción sin necesidad de hardware especializado.

4.1.2. Oportunidad de Mercado

Según el artículo de Sound Market "Visuals: su importancia en las actuaciones vivo" las visuales se han convertido en un elemento fundamental en este tipo de conciertos siendo los VJs, los técnicos encargados de su manipulación, los nuevos artistas. Este fenómeno, impulsado por redes sociales, hace que la innovación en esta tecnología sea muy atractiva. Además, según un informe de Market Research Intellect, "el mercado de la iluminación escénica programable está preparado para un crecimiento explosivo, impulsado por los avances tecnológicos y la creciente demanda de experiencias inmersivas en eventos en vivo", un mercado que va de la mano con el de las visuales. Paralelamente, el mercado de la música electrónica también ha mostrado un crecimiento significativo, con un valor estimado de 7.800 millones de dólares en 2022, proyectándose una tasa anual compuesta de alrededor del 8,2% entre 2023 y 2030.

4.2. OBJETIVOS

Los principales objetivos de este proyecto son los siguientes:

- Desarrollar tres modelos de inteligencia artificial capaces de clasificar la música por emociones, género y sección de la canción. El objetivo es que el modelo de predicción de emociones y el de identificación de sección de la canción realicen sus predicciones en tiempo real en un rango de 3-10 segundos y así que puedan responder a posibles variaciones de la música realizadas por el artista en directo. Por otro lado, para el modelo de detección de género solo se necesita que realice su predicción una vez por canción.
- Optimizar la eficiencia de los modelos mediante estrategias de preprocesamiento y ajuste de arquitectura, asegurando que puedan operar en hardware accesible sin comprometer la calidad de la predicción en tiempo real.
- Desarrollar un modelo de análisis de audio e integrarlo para la sincronización de visuales en tiempo real. Este modelo será capaz de dar reactividad básica a las visuales con el audio.
- Diseñar y construir un sistema de visualización adaptable en Touchdesigner que responda de manera fluida y variada a las predicciones de los modelos, permitiendo una experiencia visual única en cada show. Para ello, se optimizará la comunicación mediante el protocolo OSC para garantizar una baja latencia y una integración eficiente.
- Proponer un software accesible, funcional e innovador**,** que permita a artistas sin conocimientos técnicos crear experiencias visuales interactivas en tiempo real**,** sin la necesidad de grandes equipos o presupuestos.
- Explorar e innovar en el uso de Deep Learning aplicado a audio, contribuyendo al conocimiento en el área de visuales generativas y modelos de IA aplicados a la música en directo.

4.3. METODOLOGÍA

Para alcanzar los objetivos planteados, se ha seguido una metodología estructurada en varias fases. En primer lugar, se ha dedicado un tiempo considerable al planteamiento inicial del proyecto, incluyendo el análisis del problema, la definición de requisitos y el estudio en

profundidad de las tecnologías implicadas. Esta etapa ha sido clave para establecer una base sólida sobre la que construir el desarrollo posterior.

A continuación, se inició la fase de implementación, centrada en el desarrollo del código y la construcción de los modelos necesarios. Esta fase ha permitido validar progresivamente el funcionamiento de los distintos componentes.

Finalmente, en los últimos meses del proyecto, se ha enfocado el trabajo en la integración de las tecnologías utilizadas, así como en la elaboración de la memoria final, asegurando una documentación clara y completa del trabajo realizado.

4.4. PLANIFICACIÓN Y ESTIMACIÓN ECONÓMICA

Para la consecución del proyecto se va a establecer un cronograma de trabajos para poder obtener el producto final en el tiempo estimado.

Por otro lado, también se va a realizar un estudio económico del coste del proyecto con el objetivo de valorar un posible precio de reposición del producto final.

4.4.1. Planificación temporal (Cronograma)

PRIMERA FASE 1

Septiembre	**BÚSQUEDA DE PROYECTO**
Octubre	**INVESTIGACIÓN Y APRENDIZAJE** Investigación sobre deep learning y redes neuronales, técnicas como CNN y LSTM . Aprendizaje y estudio de Pytorch. Búsqueda de articulos o papers interesantes y elaboración de una estrategia inicial para el proyecto.
Noviembre	**RECOPILACIÓN DE DATOS** Recopilación de datos para los tres modelos necesarios. Descarga de audio, procesamiento y generación de espectogramas para la creación de los diferentes dataset.
Diciembre, Enero y Febrero	**ELABORACIÓN Y ENTRENAMIENTO DE MODELOS** Creación de tres modelos de inteligencia Artificial para atacar tres aspectos claves de la múscia: género, momento de la canción y emociones. Elaboración de un modelo tradicional basado en audio.

SEGUNDA FASE 1

Marzo	**CREACIÓN DE BASE EN TOUCHDESIGNER** Elaboración de un proyecto base en touchdesigner para la futura integración de los modelos
Abril	**INTEGRACIÓN DE MODELOS EN TOUCHDESIGNER** Integrar los modelos para la generación de visuales en mi base de touchdesigner. Solventar posibles problemas de latencia y sincronización
Mayo	**PRESENTACIÓN, INFORME Y MARGEN POSIBLES PROBLEMAS** Elaborar la presentación, terminar el informe y solventar posibles problemas de fases anterirores

4.4.2. Estimación Económica

Para el estudio de la estimación económica se establecerán tres subapartados:

- Costes de infraestructura y Software
- Costes de Hardware
- Costes de Desarrollo

4.4.2.1. Costes de infraestructura y Software

Recurso	Costo	Notas
Google Colab pro (2 meses)	20 €	Entrenamiento de modelos en la nube
Touchdesigner	0 €	Versión gratuita estudiantes
Almacenamiento de datos en la nube Microsoft OneDrive	20 €	Pagado con el plan de la Universidad.
Electricidad, conexión	20 €	Coste estimado.

Tabla 1. Costes de infraestructuras y software

Para el tipo de proyecto que estamos realizando, la subscripción Pro a Google Colab es suficiente para entrenar los modelos. Si este proyecto se quisiera escalar a nivel profesional el coste de entrenamiento y de hardware aumentaría significativamente, siendo necesario utilizar otros servicios y entornos más profesionalizados.

Subtotal ~ 60,00 €

4.4.2.2. Costes de Hardware

Recurso	Costo	Notas
Pc Personal	100 €	Estimación respecto al tiempo de uso

Tabla 2. Costes de Hardware

Subtotal ~ 100,00€

4.4.2.3. Costes de Desarrollo

Estimamos unas 11 horas semanales durante 8 meses, es decir un total de unas 380 horas.

Un salario de referencia de un desarrollador junior: 11 €/hora.

Recurso	Horas Estimadas	Total
Investigación y planificación	55 h	605€
Procesamiento y dataset	75 h	825€
Desarrollo de modelos IA	110 h	1210€
Integración Touchdesigner	65 h	715€
Pruebas y optimización	50 h	550€
Documentación y redacción del TFG	25 h	275€

Tabla 3. Costes de desarrollo

Subtotal ~ 4.180,00€

Coste Total del Proyecto:

Recurso	Coste Total	Porcentaje
Infraestructura y Software	60 €	1,3%
Hardware	100 €	2,3%
Desarrollo	4.180 €	96,4%

Tabla 4. Coste de proyecto

COSTE TOTAL ~ 4.340,00 €

Capítulo 5. SISTEMA/MODELO DESARROLLADO

En esta sección se describirá el sistema desarrollado y los pasos seguidos para realizarlo. Como se ha mencionado anteriormente se busca realizar un software que integre modelos de inteligencia artificial para generar visuales a través del software Touchdesigner. Para una correcta explicación del sistema se dividirá este apartado en:

- **Arquitectura de proyecto**, donde se introduce el sistema, su funcionamiento y como los modelos interactúan entre sí.
- **Recopilación y procesamiento de datos**, donde se explica cómo se han obtenido y creado los datos.
- **Creación de modelos**, donde se explica cómo se han generado los modelos de inteligencia artificial, y el modelo de análisis de audio.
- **Desarrollo del sistema de software**, donde se explica la integración de los modelos para la generación de visuales en Touchdesigner.

5.1. ARQUITECTURA DEL PROYECTO

Nuestro software se basa en predicción sobre audio. Es decir, analizamos canciones mediante técnicas de Deep Learning y análisis de audio para producir predicciones e información que luego utilizamos para generar las visuales en Touchdesigner. El sistema/software está dividido en 2 grandes partes:

1. **Configuración de la visual pre-reproducción:** por un lado, se dispone de un modelo de inteligencia artificial que predice el género de la canción con un vector de probabilidades de cada género y en función de éstas decide qué base de visuales asignar a la canción. Gracias a estas probabilidades se hace que cada visual sea distinta, pero respetando una coherencia con su género principal, es decir, dos canciones con el mismo género tendrán visuales parecidas, aunque luego se vean

afectadas por los otros géneros. Este modelo establece una visual para cada canción antes de reproducirla, de ahí que no sea un modelo de predicción en tiempo real.

2. **Configuración de la visual en tiempo real:** Por otro lado, se han generado modelos que interactúan y predicen en tiempo real para poder mover la visual durante el concierto y en función de cómo el Dj manipule la música:

    - **Modelo Predicción de emociones:** Como ya se introdujo anteriormente se ha creado un modelo capaz de predecir las emociones de las canciones en tiempo real con el objetivo de que las visuales reaccionen a estas cambiando su brillo y sus colores.

    - **Modelo Predicción de Sección:** Se ha generado también un modelo capaz de detectar el momento en el que se encuentra la canción para poder así preparar a las visuales ante posibles cambios de energía o transiciones. Por ejemplo, hacerlas vibrar si detectamos que la canción va a llegar al estribillo.

    - **Movimiento de la visual:** Por último, se dispone de un sistema de audio que hace reaccionar a las visuales de forma continua ante cambios energía, frecuencias, golpes de bombo, ritmo etc.

Gracias a la cooperación de todos los modelos entre sí, tanto los pre-reproducción como los de tiempo real, se es capaz de automatizar y mejorar la generación de visuales en vivo.

En el siguiente esquema se observa lo explicado previamente y como los modelos colaboran entre sí para generar la visual:

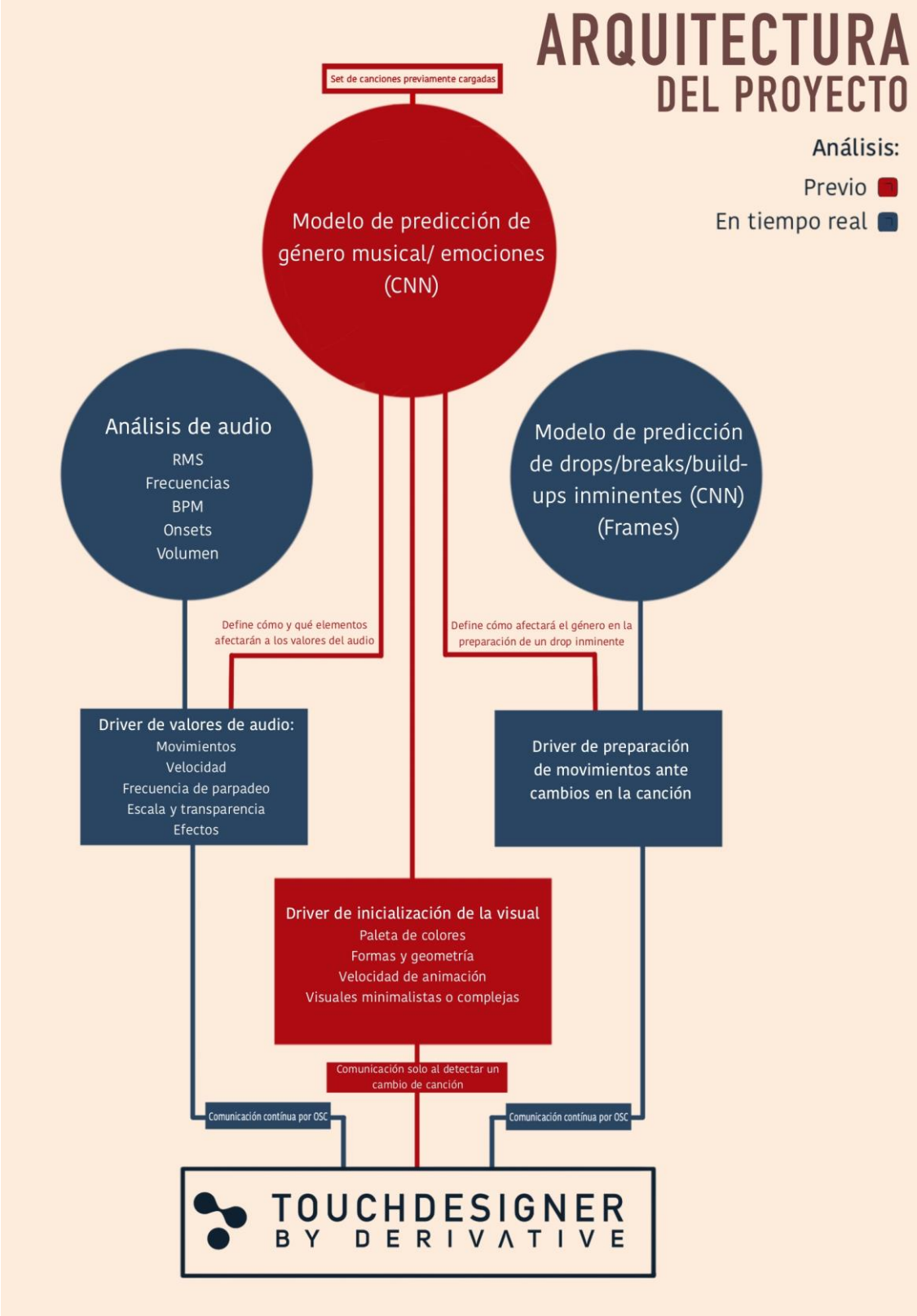

Figura 14: Arquitectura de proyecto

5.2. RECOPILACIÓN Y PREPROCESAMIENTO DE DATOS

Una de las cuestiones más importantes para abordar en este proyecto es la recopilación y generación de los datos para nuestros modelos. Se tiene la ventaja de que se cuenta con datos prácticamente infinitos y de fácil accesibilidad como son las canciones, pero al tratarse de subgéneros de música electrónica tan específicos no existen datasets clasificados para nuestro objetivo particular, por lo que tendremos que construirlos desde cero. Se comienza a continuación con una introducción sobre la extracción y forma de los datos necesarios.

5.2.1. Datos en Audio: Digitalización y Extracción de Características

El sonido es cualquier vibración que se propaga como una onda acústica a través de un medio de transmisión. Para representar estas ondas se puede utilizar una gráfica de amplitud vs tiempo, pudiendo visualizar así su comportamiento. Sin embargo, para poder analizar audio con inteligencia artificial es necesario convertir estas señales analógicas en audios digitales.

5.2.1.1. Conversión del Audio Analógico a Digital

Las señales de audio son señales continuas, es decir poseen una cantidad de valores infinitos en tiempo y amplitud. Cualquier sistema digital es incapaz de procesar datos continuos por lo que es necesario transformar esas señales continuas en señales discretas. Esto se conoce como digitalización y es necesaria cuando se quiere aplicar inteligencia artificial sobre audio. Para hacer esto existen dos procesos claves:

Muestreo (Sampling):

Como indica su nombre consiste en tomar valores o muestras de la señal en intervalos regulares de tiempo. Una mayor frecuencia de muestreo implica una mayor cantidad de muestras y por lo tanto mayor precisión respecto a la señal continua, pero también más requiere más recursos. Una tasa de muestreo muy común en audio es la de 44.1 KHz y será la que se utilice en este proyecto para la digitalización.

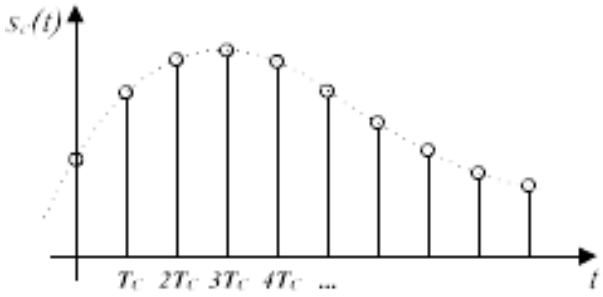

Figura 15: Ejemplo de Muestreo de señal

Fuente: https://es.wikipedia.org/wiki/Frecuencia_de_muestreo

Cuantización:

Cada muestra capturada en el proceso de muestro es asignada un valor numérico. La profundidad de bits determina la precisión con la que se representa la amplitud de sonido.

Figura 16: Cuantificación de una señal de audio

Fuente: https://es.wikipedia.org/wiki/Cuantificaci%C3%B3n_%28procesado_de_audio%29

5.2.1.2. *Representación del Audio en el dominio de la Frecuencia*

Una vez que el audio se ha digitalizado ya se puede extraer características para su análisis. Sin embargo, para nuestro objetivo es mejor representar el audio en el dominio de la frecuencia, ya que en el dominio del tiempo la información que podemos obtener es más limitada. Para esto se usa la Transformada de Fourier.

Transformada de Fourier (FT)

La Transformada de Fourier o Fourier Transform (FT) descompone una señal en una combinación de ondas sinusoidales de distintas frecuencias, permitiendo conocer la composición espectral del sonido. Sin embargo, la FT no proporciona información sobre la

variación de frecuencias en el tiempo por lo que para nuestro objetivo se necesita aplicar Short-Term Fourier Transform (STFT) que es capaz de generar un espectrograma. Un espectrograma es una representación tridimensional en tiempo, frecuencia y amplitud de la señal. Esta visualización es fundamental en tareas de clasificación musical ya que permite identificar patrones de energía de distintas bandas de frecuencia.

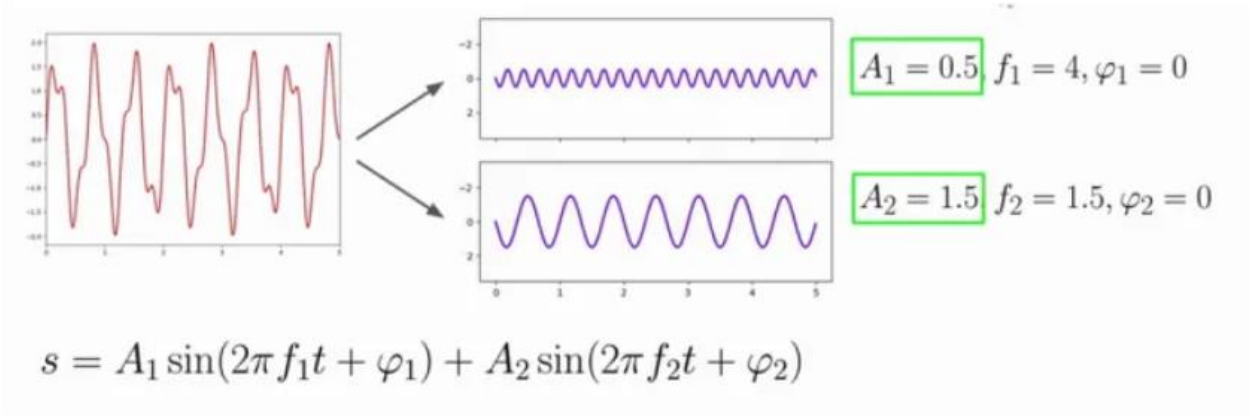

Figura 17: Ejemplo Transformada de Fourier

Fuente: http://www.sc.ehu.es/sbweb/fisica3/simbolico/fourier/fourier_1.html

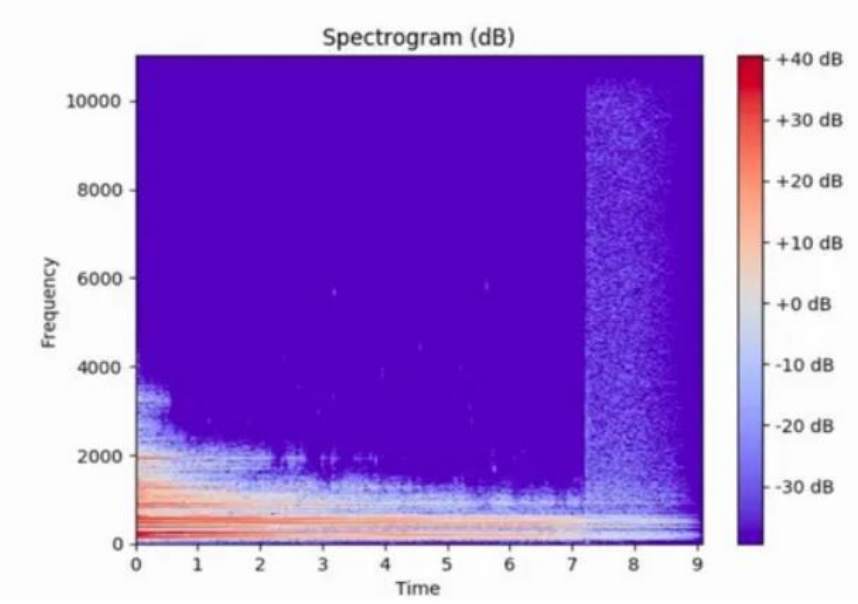

Figura 18: Ejemplo espectrograma nota de piano

Fuente: https://medium.com/@kunalkushwahatg/

5.2.1.3. Extracción de Características Avanzadas: MFCCs y Mel

El oído humano no percibe las frecuencias de la misma forma: es más sensibles a las bajas frecuencias que a las altas. Para que un modelo de IA sea capaz de capturar esto se necesita ajustar el espectrograma yendo un paso más allá. Para ello se utilizan los Coeficientes Cepstrales en la escala Mel. Estos MFCCs se generan aplicando transformaciones al espectrograma anteriormente extraído. Por una parte, se hace una conversión a la escala Mel que ajusta la representación de frecuencias para que se asemejen a cómo el oído las interpretaría y por otro se hace una transformada discreta del coseno que reduce la redundancia de los datos.

Los MFCCs son ampliamente utilizados en tareas de procesamiento de audio con Inteligencia artificial puesto que tienen más capacidad para capturar características clave del audio de manera más compacta. Se utilizan por ejemplo para reconocimiento de voz o identificación de instrumentos musicales.

Para nuestros modelos se ha decido utilizar este tipo de espectrogramas ya que como se ha explicado y también se ha mencionado en la sección de “Estado de la Cuestión” son los que mejor rendimiento dan.

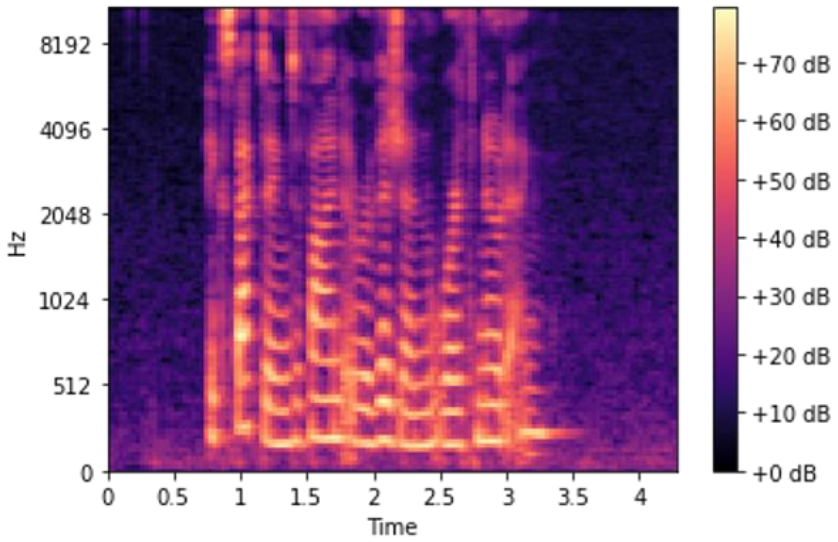

Figura 19:Ejemplo Espectrograma de Mel

Fuente: https://medium.com/@kunalkushwahatg/

5.2.2. Recopilación y Generación de datos

Una vez entendida la importancia de utilizar espectrogramas de Mel para entrenar nuestros modelos necesitamos recopilar el audio y generar los espectrogramas. En una primera fase se descargará todo el audio (canciones) y en una segunda se generarán los espectrogramas a la vez que el dataset.

5.2.2.1 Datos en Modelo de Clasificación de Género Musical

Para nuestro modelo de clasificación de género se necesita una gran variedad de canciones que abarquen los géneros seleccionados. Se ha decidido utilizar 5 de los géneros más comerciales y populares que existen hoy en día dentro de la música electrónica. Estos géneros son: Deep House, Ambient, Progressive House, Techno y Trance. Cada uno de ellos representa particularidades en términos de estructura, ritmo y elementos sonoros lo que los hace clasificables.

- ✓ **Deep House:** Se caracteriza por su ritmo relajado, líneas de bajo suaves y acordes melódicos profundos. Es un subgénero del house con influencias del jazz, el soul y el funk.
- ✓ **Ambient:** Un género centrado en la atmósfera y la textura sonora más que en ritmos marcados. Se compone de sonidos envolventes, pads etéreos y progresiones armónicas que buscan generar una experiencia inmersiva. Parecido al Deep House pero más envolvente.
- ✓ **Progressive House:** Con estructuras más largas y una construcción gradual de capas sonoras, este género se enfoca en la progresión armónica y la evolución del sonido, lo que lo hace ideal para sets de DJ extensos.
- ✓ **Techno:** Uno de los géneros más rítmicos y repetitivos, con un énfasis en beats contundentes, patrones hipnóticos y sonidos sintéticos. Su enfoque minimalista y su energía constante lo convierten en un pilar fundamental de la música electrónica.
- ✓ **Trance:** Se distingue por sus melodías envolventes, progresiones épicas y uso intensivo de efectos espaciales como reverberaciones y delays. Es un género diseñado para generar un estado emocional elevado en el oyente.

Una vez seleccionados los géneros, el siguiente paso es recopilar canciones que los representen de manera adecuada. Sin embargo, dado que muchos de estos estilos pertenecen a la misma familia y comparten características sonoras, no siempre es sencillo encontrar ejemplos completamente diferenciados.

Por ejemplo, el Deep House y el Ambient House presentan similitudes en ritmo, texturas y armonías, lo que puede dificultar una clasificación estricta. Además, muchas canciones combinan influencias de distintos géneros, reflejando la naturaleza híbrida y en constante evolución de la música electrónica.

Se ha puesto especial énfasis en seleccionar temas que representen de manera clara cada estilo de cara al entrenamiento del modelo. Sin embargo, cierta convergencia de géneros en las muestras no supone un problema para el propósito del proyecto. Al contrario, el objetivo es desarrollar un modelo capaz de adaptarse a canciones reales, que no siempre encajan en una única categoría y entrenar a la ia para que también pueda predecir esos matices.

En total se ha recopilado 4,3Gb de audio que representan unas 410-430 canciones. Por una cuestión de recursos computacionales se ha decidido dividir las canciones en fragmentos de 5 segundos para alimentar a la red neuronal. De esta forma se evitan problemas y además se optimizan las predicciones puesto que las imágenes contendrán más detalle. En total se han recopilado 14.322 fragmentos de 5 segundos.

5.2.2.1.1. Generación del Dataset

Se ha generado un csv mediante un código de Python para automatizar la creación y evitar realizarlo manualmente muestra por muestra. Este código divide cada canción en fragmentos de 5 segundos, extrae el espectrograma de Mel de cada fragmento además de características extras de la canción y clasifica cada fragmento en función de su género según la carpeta donde se encuentre. Las características extras son datos de cada fragmento que se han considerado útiles para el entrenamiento del modelo. Estas características son complementarias a la información que nos da el espectrograma de Mel e incluyen:

1. **Características de dominio temporal:**
    - RMS (Root Mean Square): Mide la energía promedio de la señal, útil para evaluar la intensidad del audio.
    - ZCR (Zero-Crossing Rate): Cuenta cuántas veces la señal cruza el eje cero, relacionado con la aspereza del sonido.
    - Mean Absolute Amplitude: Promedio del valor absoluto de la amplitud, indica el volumen medio.
    - Crest Factor: Relación entre el pico máximo y el RMS, mide la dinámica de la señal.
    - Standard Deviation of Amplitude: Variabilidad de la amplitud, útil para detectar cambios bruscos en la señal.
2. **Características espectrales:**
    - Spectral Centroid: Centro de masa del espectro, indica si el sonido es grave o agudo.
    - Spectral Bandwidth: Ancho del espectro de frecuencias, mide la dispersión del contenido armónico.
    - Spectral Roll-off: Frecuencia donde cae un porcentaje de la energía total, distingue entre ruido y tonos puros.
    - Spectral Flux: Cambio en la distribución espectral entre cuadros de análisis, útil para detectar transiciones.
    - VAD (Voice Activity Detection): Detecta presencia de voz en la señal.
    - Spectral Variation: Medida de la diferencia entre espectros consecutivos, útil para identificar variaciones en timbre.
    - Tempo: Estimación de la velocidad de la música en BPM (beats por minuto)..

5.2.2.2 Datos en Modelo de Clasificación de emociones y secciones

En este caso reutilizaremos las mismas canciones tanto para el modelo de secciones como para el de emociones.

5.2.2.2.1 Emociones

Para entrenar el modelo de clasificación de emociones, se ha utilizado una representación basada en valencia y arousal, dos dimensiones ampliamente usadas en estudios de percepción emocional en música como ya se explicó en estado del arte.

La valencia representa el grado de positividad o negatividad de una emoción, mientras que el arousal mide el nivel de activación o energía. Para cuantificarlas, se ha dividido cada dimensión en intervalos de 0.1 en 0.1, abarcando un rango de 0 a 1. Luego, se ha aplicado one-hot encoding a estas categorías, lo que permite representar todo el espectro emocional posible de manera estructurada y sin introducir relaciones arbitrarias entre clases.

5.2.2.2.2 Secciones

En el caso del modelo de clasificación de secciones, cada canción se ha segmentado en tres partes fundamentales de la música electrónica:

- **Break**: Secciones más tranquilas con menor energía.
- **Build-up**: Transiciones que aumentan la tensión antes del clímax.
- **Drop:** Secciones de mayor intensidad rítmica y energética.

Cada fragmento de audio ha sido clasificado según la sección a la que pertenece, permitiendo al modelo aprender las características específicas de cada parte.

Tanto para emociones como para secciones la clasificación ha sido manual porque no se ha encontrado una forma más sencilla de poder hacerlo. Esto implica que puede haber ciertos errores humanos o incluso ciertos sesgos a la hora de clasificar. Se ha intentado realizar la clasificación por emociones de la forma más objetiva posible siguiendo el diagrama de valencia y arousal y comparando opiniones humanas externas sobre muchas de las canciones

o géneros a las que pertenecen y las emociones que suelen transmitir. Aun así, se tiene que ser consciente de que cierta subjetividad es inevitable.

De igual forma que en la clasificación de género se han divido las canciones en fragmentos de 5 segundos y se ha extraído el espectrograma de Mel de cada uno además de las mismas características adicionales de audio. En total se han extraído unas 3.600 muestras de unas 130 canciones.

5.3. CREACIÓN DE MODELOS

5.3.1. Modelo de predicción de género musical

Como ya se ha explicado este modelo busca predecir entre 5 subgéneros de música electrónica: Ambient, Deep House, Progressive House, Trance y Techno. Para ello y basádose en los estudios ya mencionados se necesita definir una red neuronal capaz de extraer las características del audio a través de los espectrogramas y posteriormente utilizar una red recurrente que establezca relaciones entre esas características para finalmente clasificar los fragmentos en sus respectivos géneros. En este caso se ha decido juntar fragmentos de 5 segundos de tres en tres para alimentar la red. Al no necesitar predicciones en tiempo real se da a la red una ventana de tiempo más amplia para predecir, manteniendo la optimización computacional y con más datos que si solo pasásemos un fragmento de 15 segundos.

5.3.1.1 Arquitectura de la red

Como ya se explicó, se utilizan arquitecturas basadas en el estudio de X. Jiang et al. Todas ellas emplean una estructura híbrida con un primer bloque CNN y un segundo bloque RNN. Para adaptarlas a este caso se buscará optimizar la eficiencia y disminuir la complejidad de la red puesto que nuestro dataset tiene un tamaño medio. Tras probar una arquitectura CNN+LSTM se ha observado que la red mejora considerablemente utilizando una CNN+GRU y tras varias pruebas con diferentes configuraciones de esta red se han obtenido los mejores resultados con la siguiente estructura

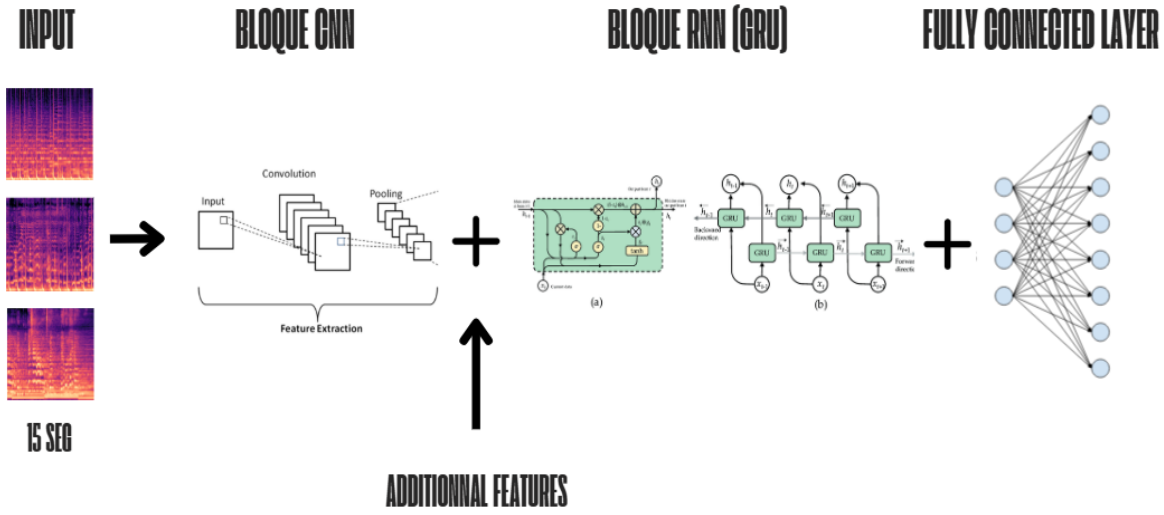

Figura 20: Arquitectura de red modelo de géneros

➢ Bloque CNN:

Un primer bloque con una red neuronal convolucional. Esta red toma imágenes de entrada con 3 canales RGB de 128x128 y las procesa por varias capas convolucionales y de pooling para extraer representaciones.

1. Conv2D (3 → 32 canales)
    - Filtro de 3×3 con padding 1 para mantener dimensiones.
    - Activación ReLU para introducir no linealidad.
    - BatchNorm para normalizar y estabilizar el entrenamiento.
    - MaxPooling (2,2) → Reduce la imagen a 64×64.
2. Conv2D (32 → 64 canales)
    - Otro bloque similar con más filtros para extraer características más complejas.
    - MaxPooling (2,2) → Reduce a 32×32.

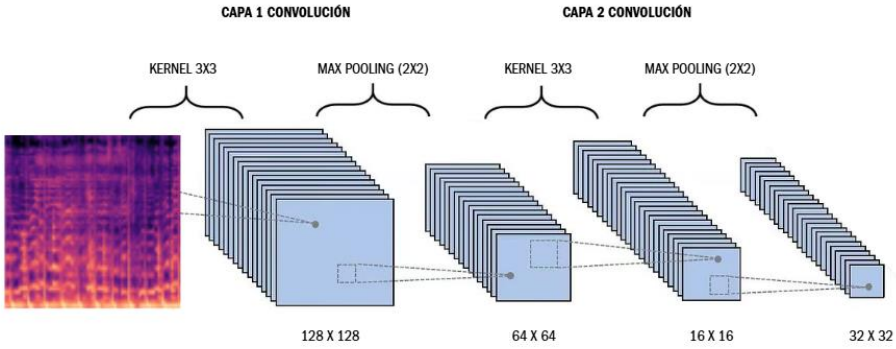

Figura 21. Esquema 1ª parte de la red convolucional

3. 3. Conv2D (64 → 128 canales)
    - Aumenta la profundidad para captar patrones más abstractos.
    - MaxPooling (2,2) → Reduce a 16×16.
4. 4. Conv2D (128 → 128 canales)
    - Última capa convolucional para refinar las características extraídas.
    - MaxPooling (2,2) → Reduce la imagen a 8×8.
5. 5. Salida del bloque CNN:

Se obtiene un tensor de características de tamaño (128,8,8), donde:

- 128 → Número de mapas de características (profundidad).
- 8x8 → Resolución espacial final después del pooling.

Este tensor se aplana en un vector de tamaño 128×8=8192 antes de pasar al siguiente bloque.

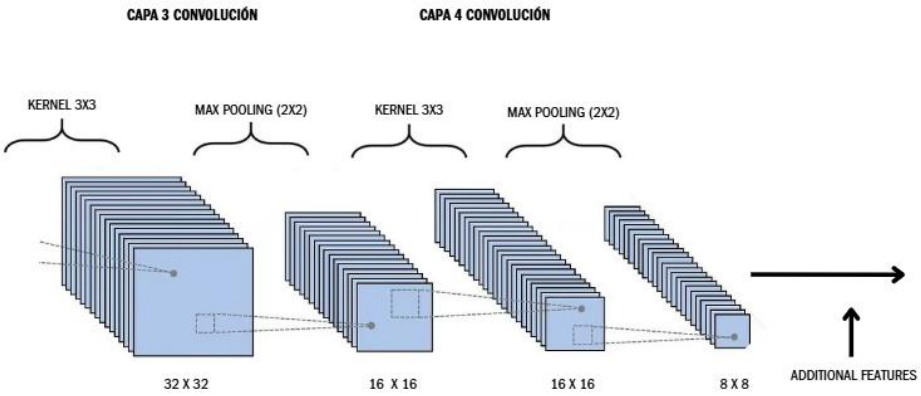

Figura 22. Esquema 2ª parte de la red convolucional

➢ Bloque RNN:

El objetivo del bloque recurrente (GRU) es procesar los datos en una secuencia y captar relaciones temporales entre ellos.

1. Entrada de la RNN:
    - Se toma la salida de la CNN (8192 características por imagen) y se concatena con las características adicionales.

2. Capa GRU:

La GRU es una variante eficiente de las RNNs que mantiene memoria de secuencias de manera optimizada. Cada unidad GRU tiene dos puertas:

- Puerta de actualización: Controla cuánta información nueva se incorpora al estado.
- Puerta de reinicio: Decide cuánta memoria pasada se olvida.

Respecto a una arquitectura LSTM la GRU es más eficiente y rápida en entrenamiento. Captura dependencias de largo plazo en secuencias sin desvanecimiento del gradiente.

La fórmula a aplicar es:

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t$$

donde h_t es el estado oculto actual, z_t es la puerta de actualización y h_t es el nuevo estado candidato.

En este caso la GRU que utilizamos tiene las siguientes características:

- 2 capas con 256 neuronas en cada una.
- Bidireccional: Captura relaciones en ambas direcciones.

3. Salida de la GRU:
    - Devuelve un vector de características de tamaño 256 * 2 (porque es bidireccional).

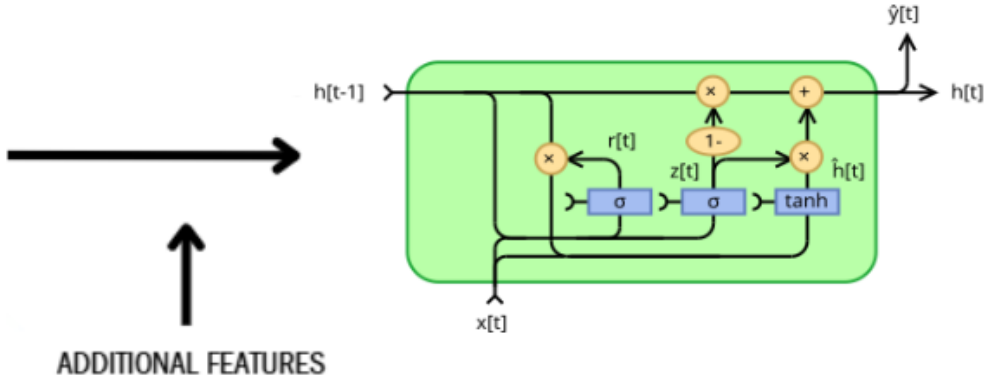

Figura 23. Esquema red GRU

➢ Capa de salida (Clasificación)

Finalmente, se usa una capa totalmente conectada para convertir la salida de la GRU en las probabilidades de clasificación.

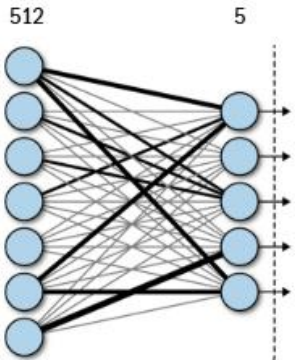

Figura 24. Esquema capa fully connected

Para entrenar la red se separan los datos en dos datasets, uno de entrenamiento con el 80% de las muestras y uno de validación con el 20%. Se ha calculado la media y la deviación estándar de todos los espectrogramas previamente para una mejor normalización. Las muestras se procesan en "batches" de 128, con tres imágenes por muestra. Para ello se ha diseñado una función de "collate" personalizada para generar el "dataloader".

Se establece como función de perdida una "CrossEntropyLoss()" para la clasificación y se configura el optimizador "AdamW", una versión mejorada del Adam con mejor regularización. Se usa un learning rate para el optimizador de 0.008. Además, se han configurado 50 épocas para el entreno con una paciencia de 5, es decir que si pasan 5 épocas sin que el modelo mejore su precisión se acaba el entreno.

5.3.1.2 Resultados modelo de género

Con la red final CNN + GRU se obtienen muy buenos resultados. La accuracy (exactitud) en entrenamiento es de 75.576% y en el caso de la validación de 64.75% Estos resultados son muy buenos ya que pertenecen a la clasificación por fragmentos de 15 segundos, no son por canción, por lo que puede haber fragmentos dentro de una canción cuya predicción es errónea, pero al final cuando hagamos la media de predicciones de fragmentos por canción la accuracy todavía será mayor.

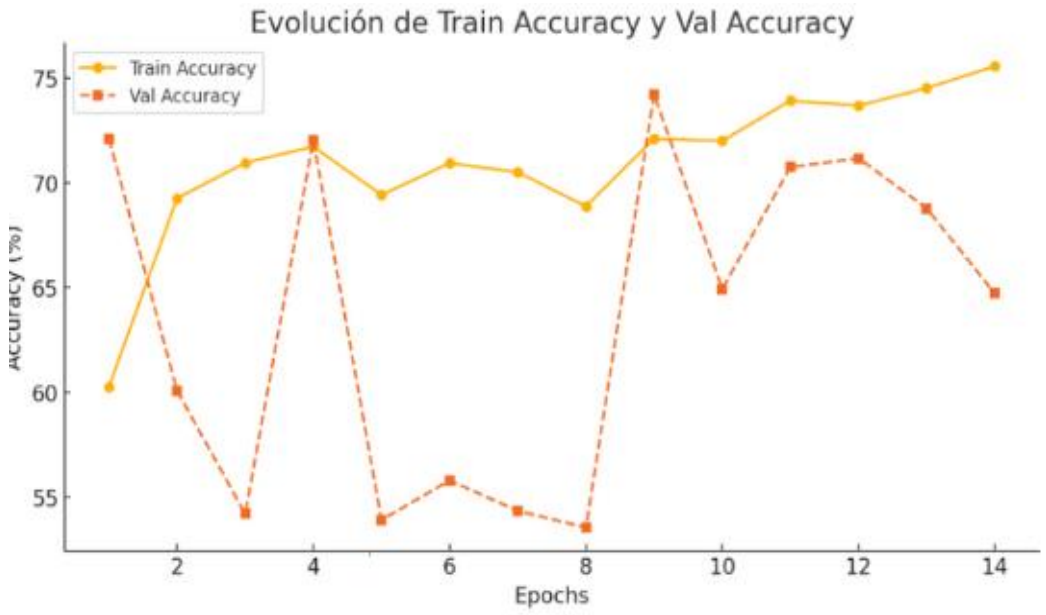

Figura 25. Evolución "Accuracy" en entreno y test 1er modelo

En el siguiente gráfico de pérdidas se puede observar como el modelo va aprendiendo haciendo cada vez más pequeña esa pérdida:

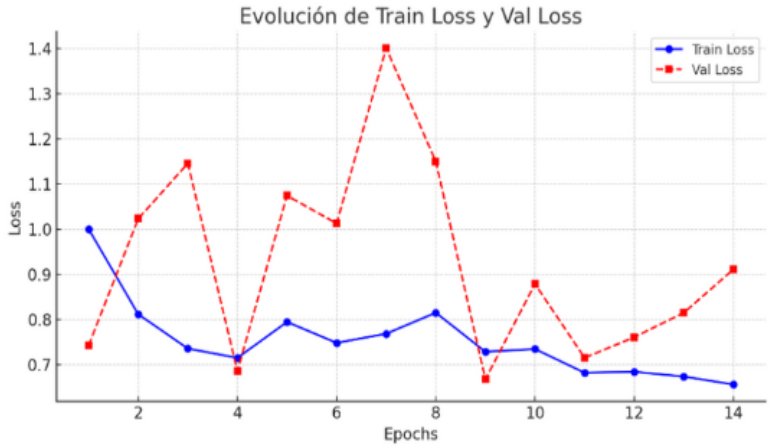

Figura 26. Evolución pérdidas en entreno y test 1er modelo

El recall o sensibilidad, que mide qué proporción de los casos positivos reales fueron correctamente identificados por el modelo, es bastante bueno en torno al 0.65 al igual que la precisión que está en torno al 0.69.

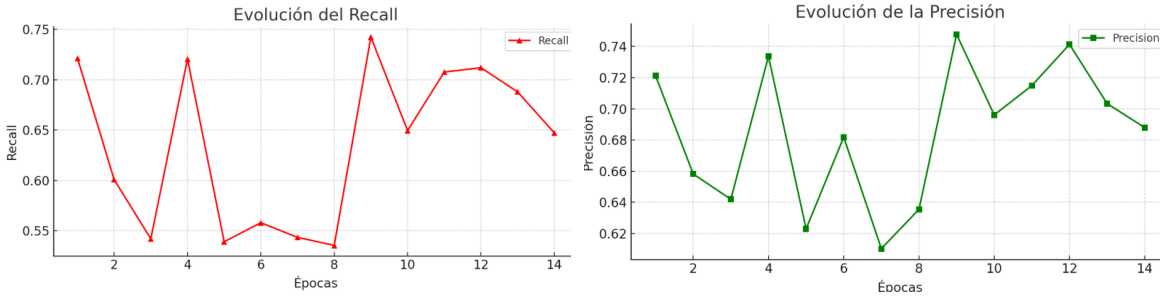

Figura 27. Evolución de la precisión y recall 1er modelo.

Como se ha explicado el conjunto de datos que se han utilizado para entrenar el modelo está compuesto por canciones altamente representativas de cada subgénero. Sin embargo, el software de visuales debe adaptarse a una gran variedad de canciones dentro de la música electrónica, muchas de las cuales combinan múltiples subgéneros y no encajan claramente en una única categoría.

Para evaluar el rendimiento del modelo en este contexto más realista, se ha generado un dataset adicional con canciones más variadas y sin una clasificación de subgénero tan

definida. Al aplicar el modelo a este nuevo conjunto de datos, los resultados de la matriz de confusión reflejan una menor precisión en las predicciones.

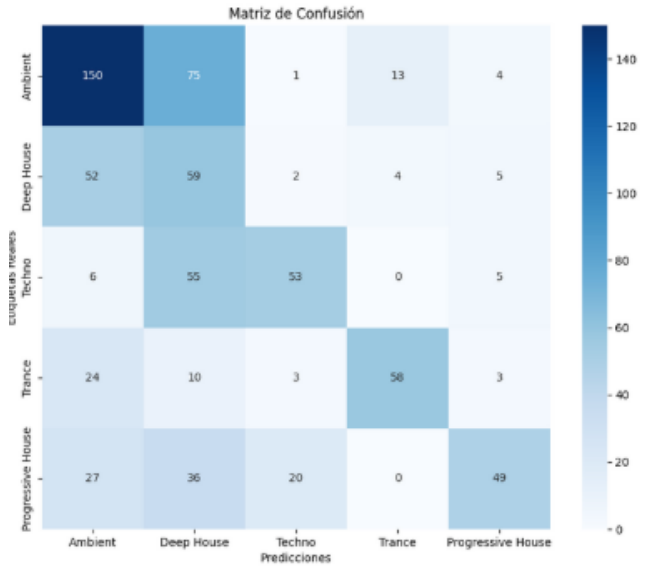

Figura 28. Matriz de Confusión 1er modelo

No obstante, el enfoque de modelado dado ofrece una ventaja clave: en lugar de realizar una clasificación rígida en un único subgénero, la red genera un vector de probabilidades que indica la afinidad de cada canción con distintos subgéneros. Este enfoque hace que el sistema diseñado sea mucho más escalable y flexible a la hora de integrarlo en las visuales.

Por ejemplo, si una canción tiene un 60% de probabilidad de ser Techno, un 25% de ser Trance y un 15% de ser Ambient, las visuales pueden configurarse para priorizar la estética Techno, pero incorporando elementos característicos de Trance y pequeños detalles de Ambient. Esto permite una adaptación dinámica de las visuales en función de la mezcla de géneros presentes en cada canción, enriqueciendo la experiencia visual de manera más natural y envolvente.

Para analizar más a fondo las predicciones del modelo con el nuevo dataset, se ha generado un archivo CSV con las probabilidades asignadas a cada canción, lo que ha permitido examinar con mayor detalle algunos de los casos clasificados erróneamente.

Al revisar los resultados, se ha observado que muchas de las predicciones incorrectas tienden a darse entre géneros similares, como Ambient y Deep House. Esto es comprensible, ya que ambos comparten características sonoras y, en algunos casos, incluso pueden ser difíciles de distinguir para un ser humano.

Asimismo, se han identificado otras muestras donde el modelo clasifica una canción en un subgénero distinto al esperado, pero con una justificación razonable. Por ejemplo, una canción etiquetada manualmente como Ambient fue clasificada por el modelo como Trance. Sin embargo, al analizar la pista en detalle, se observó que, aunque el artista pertenece al género Ambient, esa canción en particular incorporaba muchos elementos característicos de Trance.

Estos resultados refuerzan la importancia del enfoque dado en el proyecto basado en probabilidades en lugar de una clasificación rígida, ya que permite capturar la naturaleza híbrida de muchas canciones dentro de la música electrónica.

Ejemplos predicciones erróneas:

Canción	Género Real	Género Predicho	Ambient	Deep House	Techno	Techno	Trance	Progressive House
11	Ambient	Deep House	**0.392461**	**0.459741**	0.0763509	0.00259542	0.00208424	0.0667666
9	Ambient	Trance	**0.1492151**	0.1094844	0.0006803	0.0166481	**0.7173167**	0.00665
2	Techno	Ambient	**0.4120611**	0.21134	0.0632	**0.238836**	0.01724	0.05726

Tabla 5. Predicciones erróneas - género

Además, como se mencionó anteriormente se probó una arquitectura CNN + LSTM y los resultados no fueron tan buenos con un 52% de accuracy en entrenamiento y un 57% en validación:

Epoch	Train Loss	Val Loss	Train Accuracy	Val Accuracy	Val F1	Val Precision	Val Recall
1	1.323452051	1.1591327	48.86737304	0.5413642961	0.4982470442	0.4862015587	0.5413642961
2	1.179168594	1.290610877	54.67665327	0.5362844702	0.4727167011	0.5071217669	0.5362844702
3	1.096972463	1.13574584	58.23894775	0.5544267054	0.5424241046	0.5883290262	0.5544267054
4	1.074947492	1.205694115	59.11582024	0.5457184325	0.504908852	0.5360552276	0.5457184325
5	1.266789586	1.303451468	51.44318597	0.4934687954	0.4566500718	0.5122708235	0.4934687954
6	1.302024057	1.234431803	50.82206796	0.5297532656	0.4926703011	0.5289702172	0.5297532656
7	1.282288356	1.152156757	50.7124589	0.5812772134	0.5581973865	0.5857935585	0.5812772134
8	1.250208	1.148578037	52.32005846	0.5703918723	0.5516019285	0.5594723559	0.5703918723

Tabla 6. Evolución métricas modelo de género

A pesar de que la LSTM es una red más compleja y, en principio, podría parecer que ofrecerá mejores resultados, estas redes están diseñadas para capturar dependencias temporales más largas. En este caso concreto, el uso de fragmentos de 15 segundos no requiere una memoria tan extensa, lo que hace que una GRU —con una estructura más simple y menos parámetros— se adapte mejor a la tarea. Esta simplicidad no solo permite un entrenamiento más eficiente y menos propenso al sobreajuste, sino que además facilita una convergencia más rápida y estable. Estos factores pueden explicar el mejor rendimiento observado en la arquitectura CNN+GRU respecto a la CNN+LSTM.

5.3.2. Modelo de predicción de emociones

Para este segundo modelo, como ya se ha comentado, se busca predecir las emociones que transmiten las canciones. A pesar de la subjetividad de este tipo de parámetros se utiliza un sistema de clasificación de valencia y arousal que aumenta la objetividad. Una característica importante de este modelo a tener en cuenta es que las predicciones se necesitan que sean rápidas, ya que va a funcionar en tiempo real, cogiendo ventanas de 5 segundos de canción y prediciendo. La latencia por lo tanto es clave en el modelo y el diseño y arquitectura deben buscar su optimización.

Basándonos en estudios anteriores se ha empezado probando una arquitectura de red muy parecida a la del anterior modelo de género, pero en este caso utilizando regresión al tratarse de valores continuos.

El funcionamiento es muy parecido a diferencia de que en este caso se evalúa por muestras de 5 segundos en lugar de 15, y que la red recurrente utilizada es una LSTM en lugar de una

GRU sustituyendo el bloque CNN anterior por una red ya definida conocida como Resnet-18.

La Resnet-18 es una red residual desarrollada por Microsoft y cuyo número 18 hacer referencia al número de capas con pesos. En concreto la Resnet-18 destaca por ser muy ligera, ideal para predicciones rápidas, pero muy eficiente igualmente. Lo característico de una ResNet son sus "bloques residuales", que incluyen conexiones tipo "skip" o "atajos". En vez de pasar solo de una capa a la siguiente se aplica la siguiente lógica:

$$y = F(x, \{Wi\}) + x$$

Esto permite que la red aprenda la diferencia entre la entrada y la salida esperada, no la transformación entera.

La ventaja de utilizar estas redes ya definidas es que ya están probadas y son muy potentes y eficientes.

Sin embargo, tras realizar varios entrenamientos utilizando diferentes criterios de evaluación, como el Mean Absolute Error (MAE) y el Mean Squared Error (MSE), los resultados no fueron satisfactorios. Estos errores se deben al tamaño relativamente pequeño de la muestra, que no es lo suficientemente grande para un modelo de Deep Learning, y a su distribución desigual. Como consecuencia, el modelo tiende a predecir la media de todas las muestras, lo que le permite minimizar el error a tan solo 0.18. No obstante, esta aproximación no refleja una predicción precisa, ya que el modelo no está capturando correctamente las variaciones específicas de los datos:

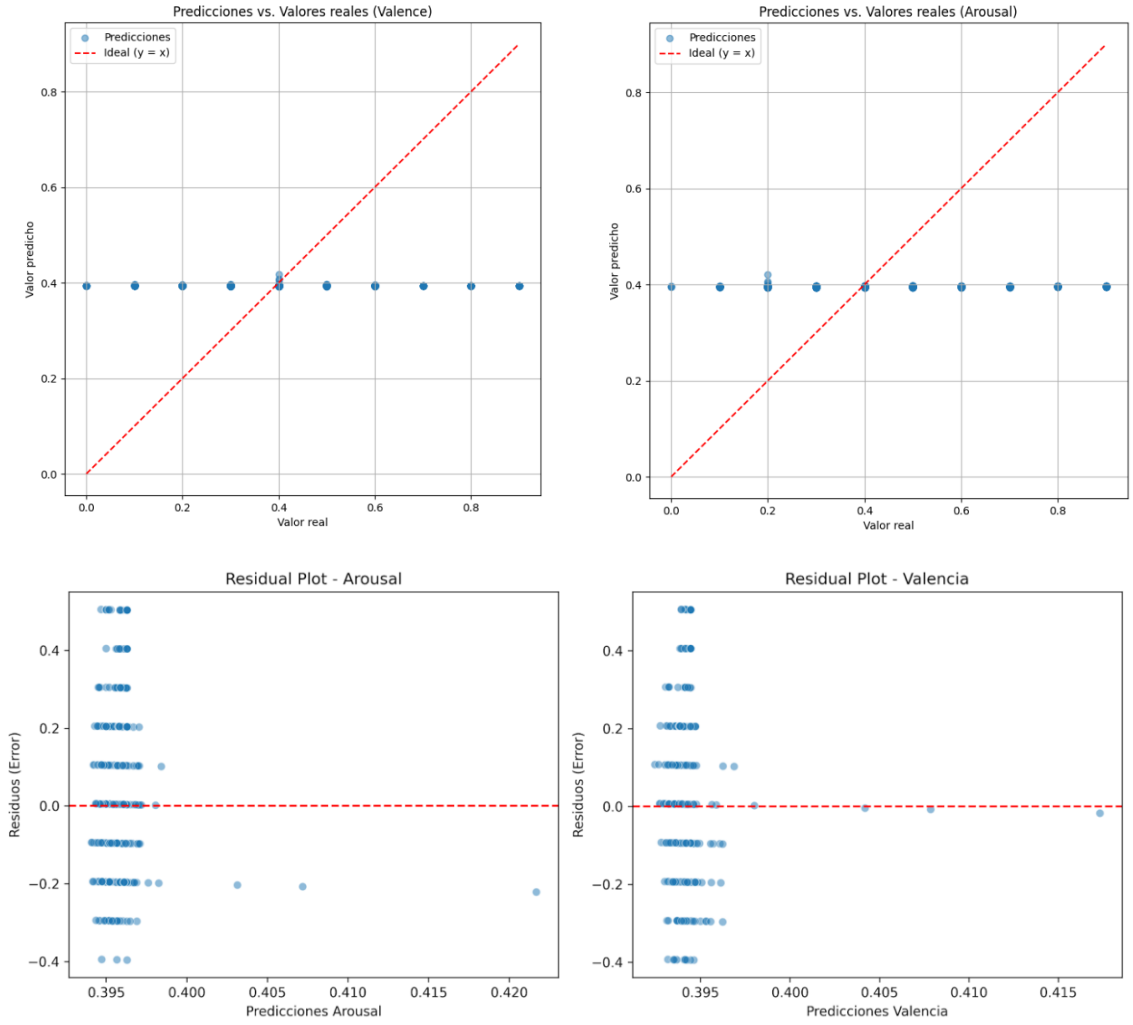

Figura 29. Resultados modelo descartado de emociones

Por lo tanto, se decidió explorar otras opciones para conseguir el objetivo. En lugar de utilizar una red neuronal recurrente para la segunda parte de la arquitectura se sustituyó dicha red por un modelo de machine learning Random Forest.

Random Forest es un algoritmo de aprendizaje supervisado que combina múltiples árboles de decisión para realizar predicciones más robustas y precisas. Su funcionamiento se basa en el principio de ensamblado: cada árbol vota y la predicción final se obtiene por mayoría (en clasificación) o por promedio (en regresión) que es el que usaremos en este caso. Esta técnica permite reducir el riesgo de overfitting y mejora la generalización del modelo.

Esta sustitución presenta varias ventajas frente a las redes neuronales recurrentes como menor complejidad computacional**,** mayor rapidez en el entrenamiento y mejor interpretabilidad del modelo, eso sí sacrificando la capacidad de capturar dependencias temporales y secuenciales complejas.

5.3.2.1. Arquitectura de la red

Ahora la red quedaría de la siguiente forma:

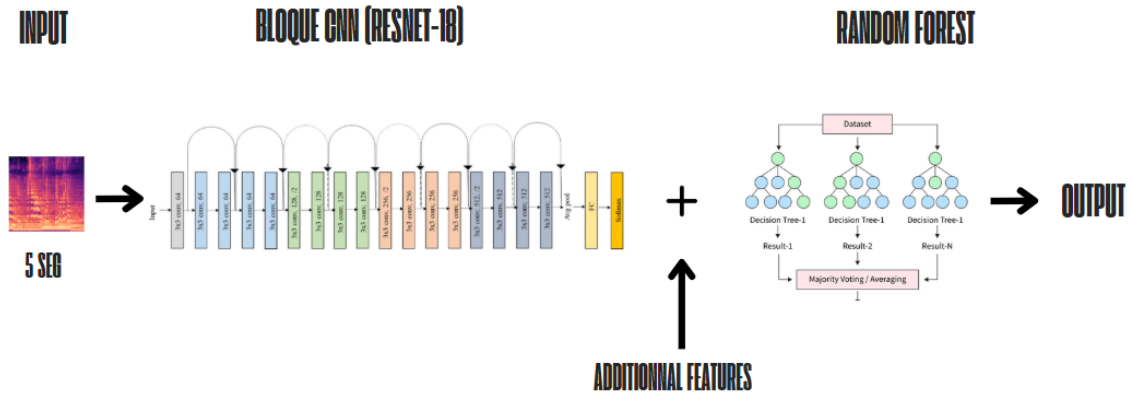

Figura 30. Arquitectura de la red modelo de emociones

El proceso de entrenamiento es un poco distinto ya que en este caso no se crea una red neuronal completa, sino que simplemente se extraen características de las imágenes y junto con las características adicionales se introducen al Random Forest que genera una predicción. En este caso se ha creado un modelo para cada salida, uno para valencia y otro para arousal.

Cada modelo está compuesto por 100 árboles de decisión, lo que permite una combinación equilibrada entre precisión y capacidad de generalización. Para el entrenamiento, se ha utilizado un tamaño de lote (batch size) de 128, lo que facilita un procesamiento eficiente de los datos mediante dataloaders personalizados. Además, se ha fijado el parámetro random_state=42 con el objetivo de asegurar la reproducibilidad de los resultados obtenidos durante el proceso de entrenamiento.

5.3.2.2. Resultados modelo de predicción de emociones

Con el trabajo desarrollado aplicando esta técnica los resultados obtenidos son los siguientes:

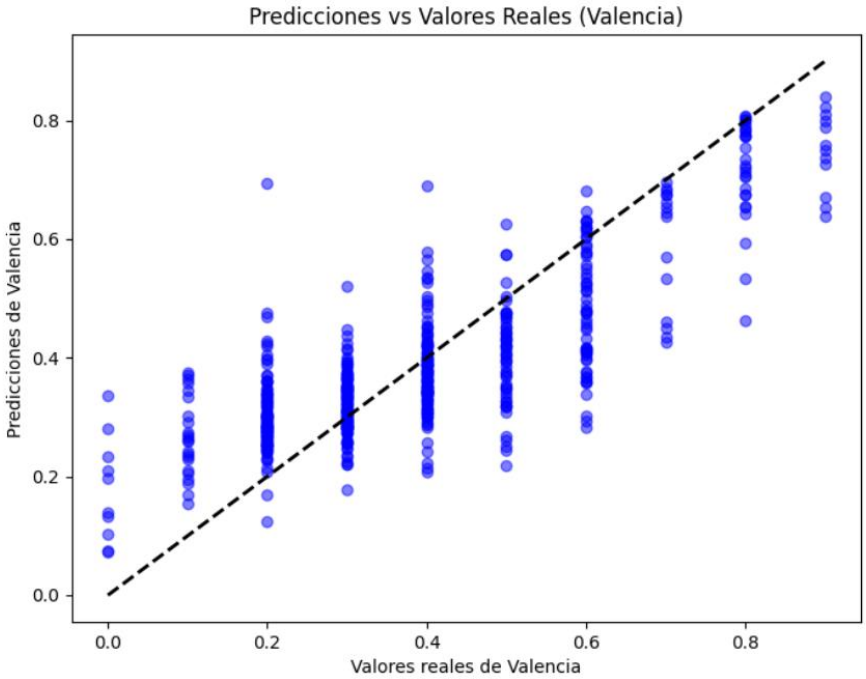

Figura 31: Resultados de Valencia

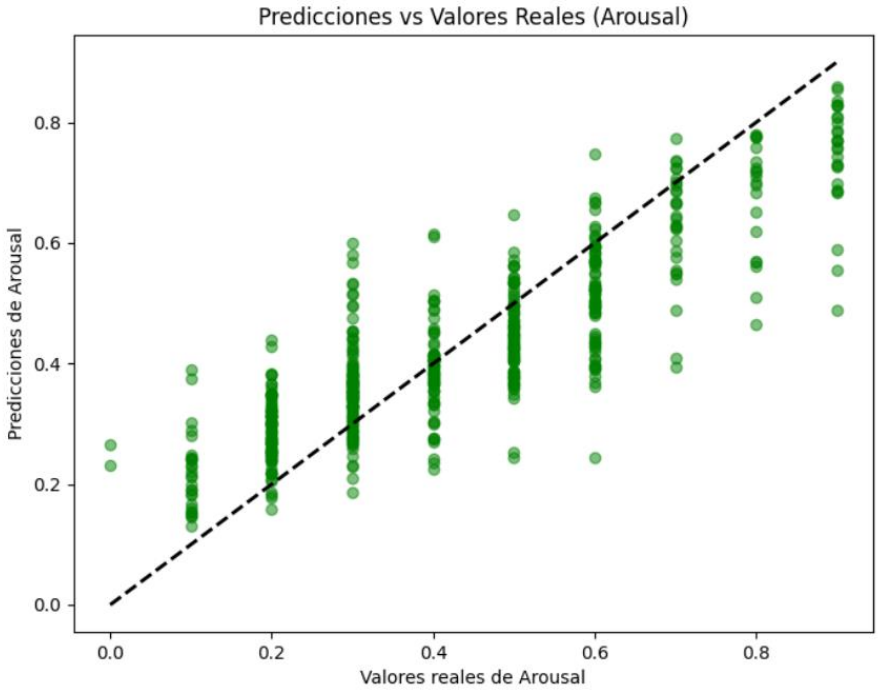

Figura 32: Resultados de Arousal

Métrica	Valor
MSE_Valencia	0.01307630209
MSE_Arousal	0.01207742584
MAE_Valencia	0.08798703733
MAE_Arousal	0.08660370362
R2_Valencia	0.6564977935
R2_Arousal	0.7234392251

Tabla 7. Métricas modelo de emociones

Los resultados obtenidos por el modelo de Random Forest muestran un rendimiento sólido. El error cuadrático medio (MSE) y el error absoluto medio (MAE) son bajos tanto para valencia como para arousal, lo que indica que las predicciones están, en promedio, cerca de los valores reales. Además, los coeficientes de determinación (R^2) alcanzan valores de 0,657 en valencia y 0,723 en arousal, lo que significa que el modelo logra explicar un porcentaje considerable de la variabilidad presente en los datos. Dado que las tareas de predicción emocional suelen implicar una alta subjetividad y ruido en las etiquetas, estos resultados se consideran razonablemente buenos.

5.3.3 Modelo de predicción de secciones

Este modelo constituye el más sencillo de los tres desarrollados, en cuanto al número de clases a predecir, ya que la red neuronal debe clasificar el fragmento de audio en una de las tres posibles secciones de una canción electrónica: break, pre-drop y drop. A pesar de su simplicidad estructural, este modelo tiene una relevancia clave dentro del sistema, ya que la predicción de la clase pre-drop es especialmente importante. Esta sección actúa como un anticipo del clímax musical y nos proporciona una ventana de tiempo ideal para preparar los efectos visuales que acompañarán el drop, permitiendo generar transiciones más impactantes y espectaculares en la experiencia audiovisual.

Para abordar esta tarea, se ha seguido una estrategia de arquitectura similar a la empleada en el resto de modelos desarrollados: una red neuronal que combina una primera capa

convolucional (CNN), encargada de extraer características relevantes del espectrograma de audio, seguida de una red neuronal recurrente (RNN) que interpreta la secuencia temporal de dichas características. En este caso, se ha optado por emplear una arquitectura LSTM (Long Short-Term Memory) como variante de RNN. Esta elección se debe a la necesidad de mejorar la capacidad del modelo para detectar patrones temporales complejos y mantener información relevante a lo largo de la secuencia. Las LSTM han demostrado un rendimiento superior respecto a las RNN tradicionales en tareas donde el contexto y la evolución temporal del dato son determinantes, como es el caso del análisis de secciones musicales.

Aunque este modelo, al igual que el de emociones, realiza predicciones cada cinco segundos, en este caso se ha decidido priorizar la precisión sobre la velocidad. Esto se debe a que, en el contexto de música electrónica, la sección pre-drop suele extenderse durante un intervalo de al menos 5 a 15 segundos, lo que proporciona un margen temporal suficiente para actuar visualmente sobre la escena. Por tanto, aunque el modelo no funcione en tiempo real estricto, sí permite realizar una integración efectiva y coherente con el flujo audiovisual del espectáculo.

5.3.3.1. Arquitectura de la red

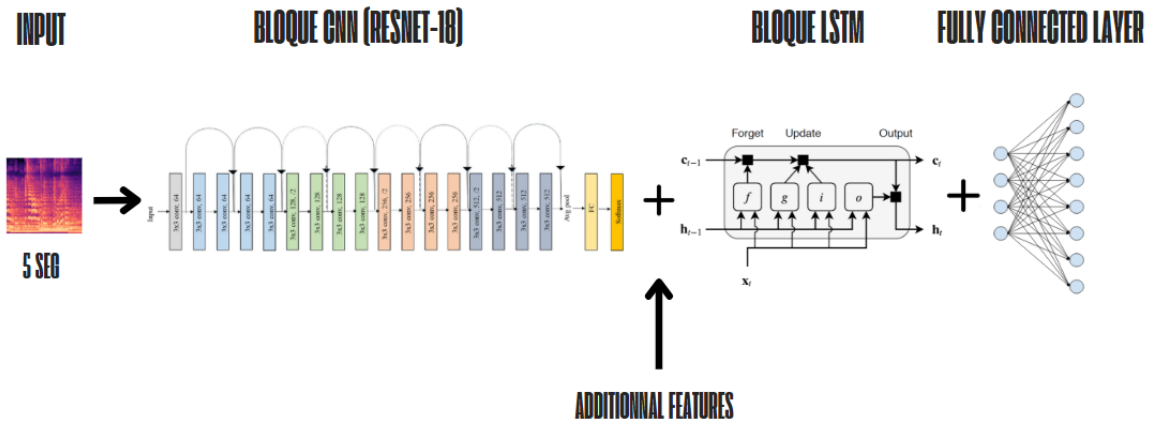

Figura 33. Arquitectura red modelo de secciones

1. **ResNet-18 (CNN):**

Se usa una ResNet-18 preentrenada sin su capa final para extraer un vector de 512 características por espectrograma.

2. **Concatenación de características:**

Este vector se concatena con otras características adicionales (como datos numéricos).

3. **LSTM bidireccional:**

Una LSTM con dos capas y bidireccional procesa la secuencia completa, capturando relaciones temporales en ambos sentidos.

4. **Capa final (FC):**

La salida de la LSTM se pasa a una capa lineal para obtener la predicción final entre las clases definidas.

Para el entrenamiento del modelo se ha utilizado un learning rate de 0.008, un peso de regularización (weight decay) de 1e-4 y un total de 50 épocas. Además del espectrograma, se han incorporado 10 características adicionales por muestra, que se concatenan al vector extraído por la red convolucional antes de ser procesado por la LSTM. Se emplea un batch size de 128 y se han definido transformaciones de normalización basadas en la media y desviación típica del conjunto. La red se entrena sobre una GPU de Google Colab, aprovechando el paralelismo para acelerar el proceso.

5.3.3.2. Resultados modelo de predicción de secciones

Epoch	Train Loss	Val Loss	Train Accuracy	Val Accuracy	Val F1	Val Precision	Val Recall
1	1.115326086	0.9429304481	0.4246124941	0.5403377111	0.3790919873	0.291964842	0.5403377111
2	0.9768908024	0.921611774	0.5063410052	0.5403377111	0.3790919873	0.291964842	0.5403377111
3	0.9631817411	0.9408560753	0.5063410052	0.5403377111	0.3790919873	0.291964842	0.5403377111
4	0.9539706952	0.9183988571	0.5063410052	0.5403377111	0.3790919873	0.291964842	0.5403377111
5	0.9408165427	0.9084435344	0.4992954439	0.5403377111	0.3790919873	0.291964842	0.5403377111
6	0.9039390438	1.429934669	0.5265382809	0.5403377111	0.3790919873	0.291964842	0.5403377111
7	0.863751443	0.9329612136	0.5988727102	0.521575985	0.4871352605	0.5322933913	0.521575985
8	0.8400612368	0.8373896599	0.6181305777	0.5834896811	0.539894361	0.5856522327	0.5834896811
9	0.8388174983	0.9400591969	0.6110850164	0.5684803002	0.4520833882	0.5565047654	0.5684803002
10	0.7910246989	1.067250383	0.635979333	0.5009380863	0.4506188452	0.6007252874	0.5009380863
11	0.7706899117	1.11311003	0.6505401597	0.3996247655	0.2964351704	0.555094685	0.3996247655
12	0.7438625553	0.725296998	0.6669798027	0.6923076923	0.641157417	0.6155751126	0.6923076923
13	0.7191161058	0.9191813111	0.6815406294	0.5328330206	0.4884135502	0.5610651883	0.5328330206
14	0.7145541591	0.6747302771	0.6810709253	0.7110694184	0.6610761689	0.6177369739	0.7110694184
15	0.6929664997	0.7083257198	0.6928135275	0.7091932458	0.6582783965	0.6285114809	0.7091932458
16	0.6726785653	1.064199674	0.6998590888	0.5966228893	0.5033475071	0.5818957467	0.5966228893
17	0.679206182	0.6927170157	0.6979802724	0.6885553471	0.6380741197	0.613027395	0.6885553471
18	0.651578223	0.926563704	0.7130108032	0.6153846154	0.5761508615	0.6044543547	0.6153846154
19	0.6342661907	0.6356309056	0.7205260686	0.6923076923	0.6791779856	0.6714532846	0.6923076923
20	0.6213558912	0.654818666	0.7266322217	0.7166979362	0.7176711328	0.7226758232	0.7166979362
21	0.6475646005	0.7052781463	0.7134805073	0.6923076923	0.6827112995	0.6797571554	0.6923076923
22	0.6236693228	0.6337984562	0.7397839361	0.7091932458	0.6702954707	0.7058591075	0.7091932458
23	0.6213355117	0.7827945948	0.7299201503	0.6566604128	0.6653312382	0.681415359	0.6566604128
24	0.6153573639	0.9387979746	0.7421324566	0.6116322702	0.5768013824	0.6885445022	0.6116322702
25	0.5891281331	0.6737963617	0.7595115078	0.6754221388	0.6634567038	0.6961078706	0.6754221388
26	0.5930507709	0.5847488821	0.7407233443	0.7298311445	0.6910026517	0.6974622843	0.7298311445
27	0.5627806678	0.5450807273	0.7689055895	0.7448405253	0.7408338079	0.7389645046	0.7448405253
28	0.5720510097	0.6816696286	0.7581023955	0.6848030019	0.6828929537	0.7558307121	0.6848030019
29	0.5353050162	0.5564684749	0.7782996712	0.7504690432	0.7328966698	0.7502446714	0.7504690432
30	0.5269504996	0.5955817878	0.7839361202	0.7317073171	0.697461294	0.721546649	0.7317073171
31	0.5118576639	1.323448491	0.7975575388	0.5853658537	0.5686804077	0.7177668344	0.5853658537
32	0.5116551536	0.9093252659	0.7923907938	0.6904315197	0.6735016961	0.7430787024	0.6904315197

Tabla 8. Métricas modelo de secciones

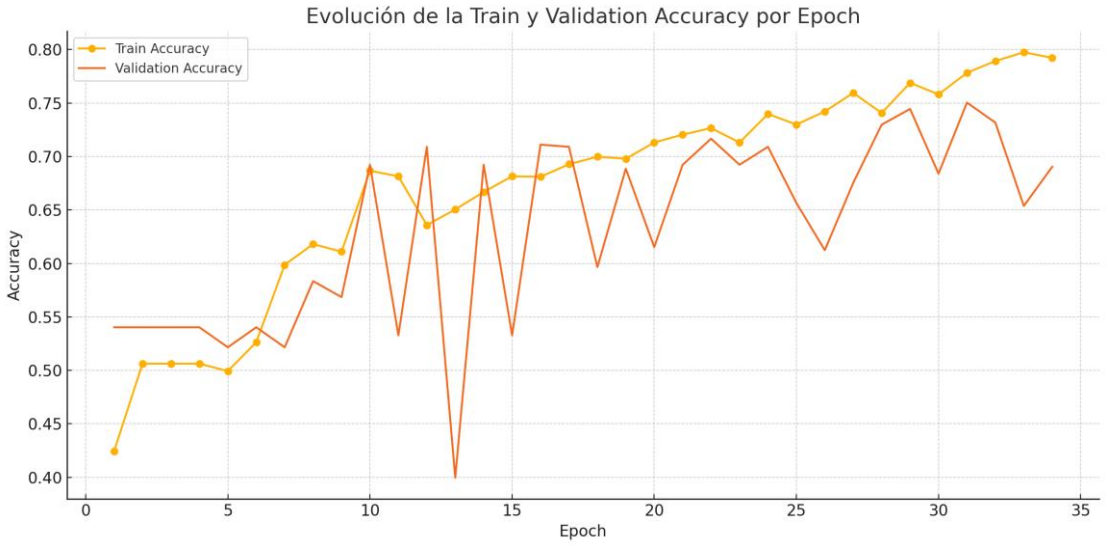

Figura 34. Evolución "Accuracy" entreno y test modelo de secciones

El modelo alcanza una precisión del 79,2 % en el conjunto de entrenamiento y del 69 % en el conjunto de test. A primera vista, estos resultados pueden parecer razonablemente buenos, a pesar de que se trata de una tarea de clasificación con tres clases posibles, lo que estadísticamente facilita al modelo obtener una predicción correcta en comparación con escenarios de clasificación más amplios. Sin embargo, un análisis más detallado revela ciertas debilidades importantes en su rendimiento.

Uno de los principales problemas detectados es la distribución desequilibrada de las clases en el conjunto de datos. Dos de las tres clases cuentan con una cantidad significativamente mayor de muestras que la tercera, lo que genera un sesgo en el entrenamiento del modelo. Como consecuencia, tiende a clasificar las instancias dentro de estas clases mayoritarias, ignorando en muchos casos la clase minoritaria, incluso cuando sería la predicción correcta. Este fenómeno, conocido como bias de clase o class imbalance, afecta directamente a la capacidad del modelo para generalizar de forma justa y equilibrada.

La causa principal de este desequilibrio radica en el número reducido de muestras disponibles para el entrenamiento. Al tratarse de un conjunto limitado de ejemplos, no se ha podido representar de forma adecuada la diversidad de situaciones y patrones que podrían encontrarse en datos reales. Esta falta de variabilidad limita el aprendizaje del modelo y reduce su capacidad para identificar correctamente las características distintivas de la clase con menos representaciones.

En consecuencia, aunque la métrica global de accuracy parece aceptable, no refleja adecuadamente la calidad del modelo en contextos prácticos donde una distribución equilibrada o una buena sensibilidad en todas las clases resulta crítica. Esto pone de relieve la necesidad de aumentar la cantidad y diversidad del dataset o de aplicar técnicas específicas para mitigar el desbalance, como el sobre muestreo, la ponderación de clases o el uso de métricas adicionales como el F1-score por clase.

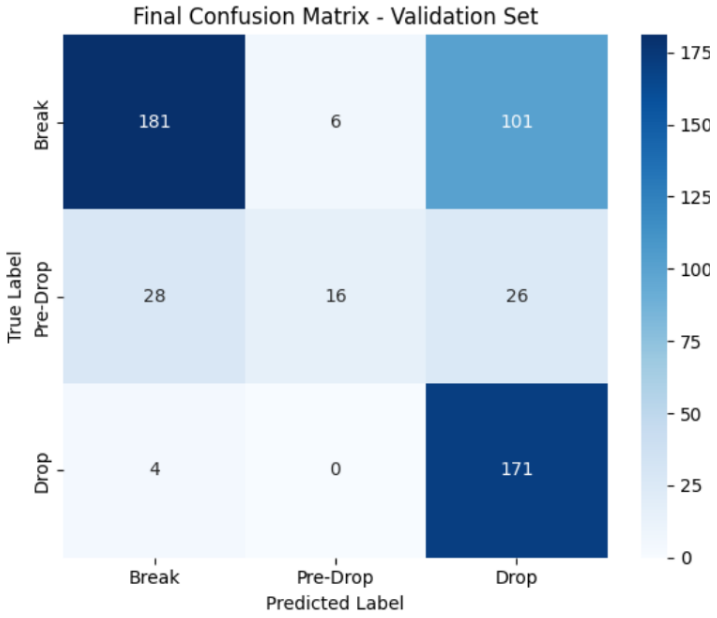

Figura 35. Matriz de confusión - secciones

Aun así, se ha decidido implementar el modelo en TouchDesigner y probar su integración utilizando únicamente aquellas predicciones que se consideren correctas o representativas. Esto permitirá evaluar el comportamiento visual en condiciones controladas. Cabe destacar que la principal limitación no reside en la arquitectura del modelo, sino en la escasez de datos disponibles para el entrenamiento, un aspecto que puede abordarse y mejorar fácilmente en futuras iteraciones del proyecto.

5.3.4. Modelo de Audio

Para el modelo de audio se han utilizado las herramientas del propio software de visuales de Touchdesigner. Al final son básicamente funciones de Python ya creadas y nos ayudan a agilizar el proceso. Como se puede observar en la figura 31, tenemos la señal de audio de entrada que, para el uso del software en vivo, vendría de la controladora del DJ. En primer lugar, se procesa la señal, como se puede ver también en la figura 31, de tres formas distintas para obtener sus diferentes rangos de frecuencias: bajas, medias y altas. Esto se consigue añadiendo filtros de paso bajo, medio y alto en función de las frecuencias que se quieren obtener. En este caso se establece el filtro de paso bajo en 180 Hz, el de paso medio en 800 Hz y el de paso alto en 3500 Hz.

Figura 36. Esquema separación de frecuencias

Una vez separado el audio en frecuencias se extraen componentes que luego se puedan mapear al movimiento de la visual

Frecuencias:

En primer lugar, se procesan cada una de las salidas según frecuencia filtradas previamente para obtener una mejor separación y válida para la integración. Se seleccionan los rangos correspondientes a graves, medios y agudos utilizando operadores Limit, que aíslan las frecuencias deseadas. Posteriormente, mediante operadores Add, se suma la energía total de cada rango para obtener un único valor representativo por banda. Estos valores se suavizan con filtros para eliminar fluctuaciones bruscas y producir una salida estable. El resultado final son tres señales ("low", "mid" y "high") que representan de manera simplificada la energía en graves, medios y agudos, y que se utilizan para animar visuales de forma más fluida y controlada.

Figura 37. Esquema análisis de frecuencias

Además de la separación por rangos de frecuencia, se extraen elementos rítmicos clave como el Kick, el Snare y el Rhythm general de la canción. El Kick y el Snare son componentes

esenciales de la percusión y su detección es fundamental para generar efectos visuales reactivos, como pulsaciones o movimientos bruscos sincronizados con la música.

Kick (bombo):

Para detectar el Kick, se parte de la señal correspondiente a las frecuencias bajas previamente separadas. Sobre esta señal se aplica un análisis de variaciones bruscas en la amplitud: cuando la energía en las frecuencias graves presenta un aumento repentino (pico), se interpreta como la presencia de un Kick. Esta técnica de detección basada en la energía permite identificar el golpe de bombo en la gran mayoría de las canciones (aproximadamente un 99% de efectividad).

Una vez detectados los Kicks, se genera una señal de trigger (disparo) que se activa cada vez que ocurre un evento de este tipo. Esta señal controla un Switch que activa o desactiva determinados comportamientos visuales en sincronía con el Kick. Además, se parametriza a través de una pequeña interfaz de usuario que permite ajustar la sensibilidad y el comportamiento del sistema, facilitando su adaptación a diferentes canciones o estilos musicales.

Snare (Caja):

De manera similar al Kick, se extrae también el Snare, otro elemento fundamental de la percusión. Mientras que el Kick se detecta principalmente en las frecuencias bajas, el Snare suele encontrarse en un rango de frecuencias medias.

Para detectar el Snare, se utiliza la señal correspondiente a las frecuencias medias previamente separadas. Se analiza esta señal en busca de variaciones bruscas de energía, de forma que cuando se produce un pico repentino en este rango, se interpreta como la presencia de un golpe de Snare. Esta aproximación resulta efectiva porque los snares, al ser golpes secos y brillantes en la percusión, generan aumentos característicos en la energía de las frecuencias medias.

Al igual que con el Kick, los eventos de detección de Snare se transforman en triggers que activan un Switch, permitiendo generar respuestas visuales específicas (como destellos,

cambios de color o transiciones rápidas) cada vez que suena un Snare. Además, se parametriza la sensibilidad mediante una interfaz de usuario, que facilita el ajuste del sistema para adaptarlo al carácter rítmico de cada canción.

Rythm (Ritmo):

Para detectar el Rhythm, se analiza la energía general del audio, buscando capturar la dinámica continua de la música más allá de eventos puntuales como el kick o el snare. Se suaviza la señal para seguir mejor las variaciones de energía en el tiempo, y se ajusta mediante filtros para enfocarse en los cambios más relevantes. Después, se define un umbral a partir del cual consideramos que hay suficiente intensidad para activar un evento de ritmo. Cada vez que la señal supera este umbral, se genera una activación que puede ser usada en las visuales, permitiendo que se muevan de forma continua y fluida siguiendo el pulso general de la canción.

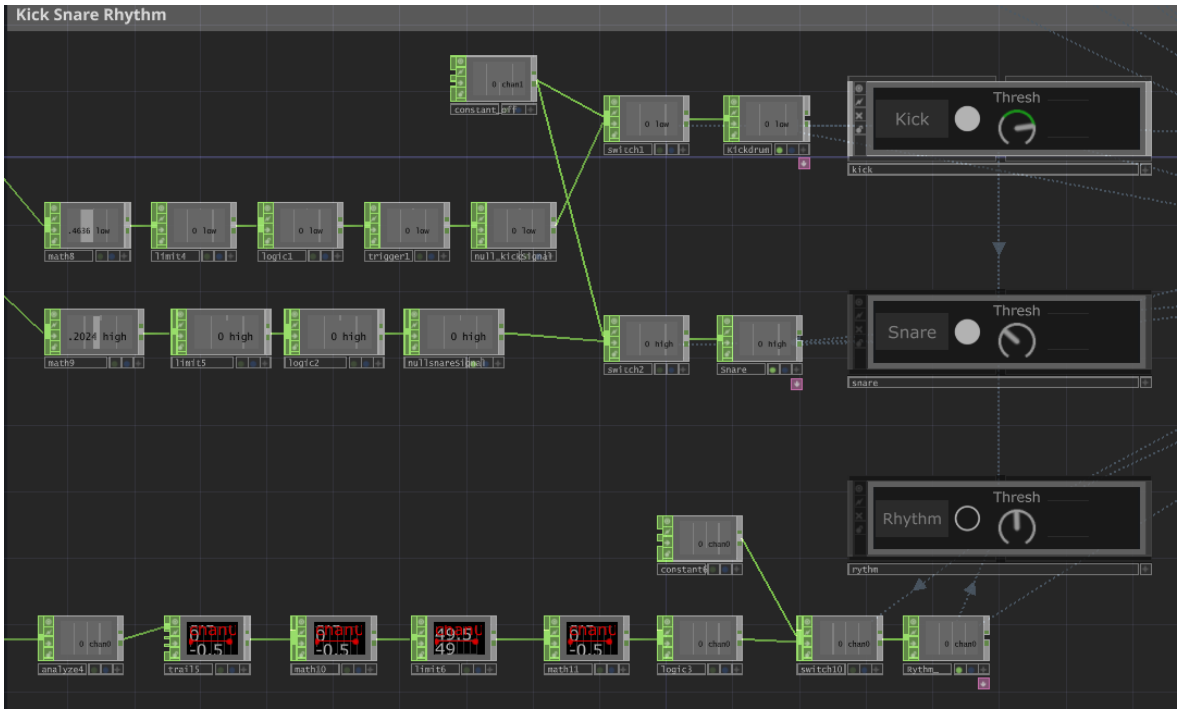

Figura 38. Esquema análisis bombo, caja y ritmo

Densidad espectral y centroide:

Por último, se quiere obtener la densidad espectral y el centroide. Para ello, se analiza el contenido en frecuencias del audio de manera más detallada.

La Spectral Density indica cuánta energía hay concentrada en diferentes zonas del espectro (bajas, medias o altas frecuencias), permitiendo saber si una parte de la canción está más cargada en graves, medios o agudos.

El Spectral Centroid, en cambio, mide el "centro de gravedad" de la energía espectral: si predomina en frecuencias bajas o altas. Es un indicador útil para capturar cambios de intensidad o "brillo" en la música.

Ambas medidas son procesadas y filtradas para estabilizar su comportamiento y adaptarlas a los parámetros que alimentarán las visuales, haciendo que reaccionen de forma más rica y sutil al contenido musical.

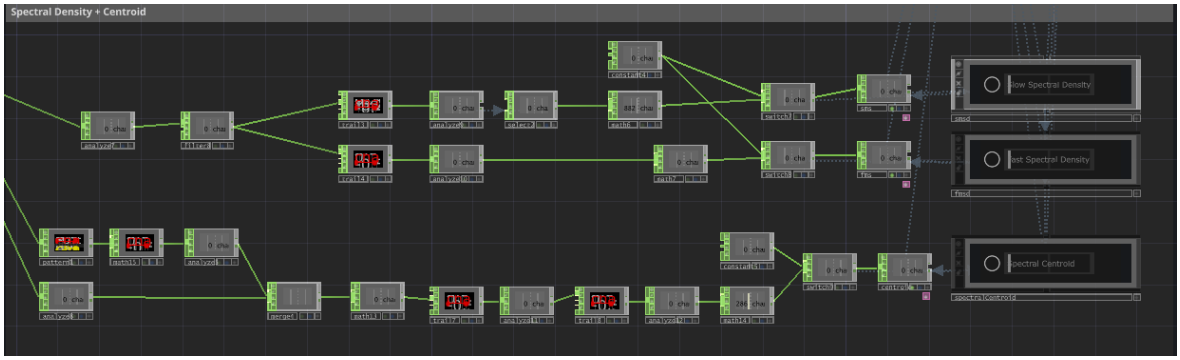

Figura 39. Esquema análisis de la densidad espectral y centroide.

5.4. INTEGRACIÓN DE MODELOS, SOFTWARE Y RESULTADOS

El último paso para finalizar el software ha consistido en la integración de los modelos desarrollados en el entorno visual de TouchDesigner, permitiendo su funcionamiento conjunto con el análisis de audio y la lógica visual en tiempo real.

Para facilitar esta integración se ha utilizado el protocolo OSC (Open Sound Control)**,** una solución de comunicación ampliamente empleada en entornos multimedia por su eficiencia, baja latencia y simplicidad de implementación. OSC permite la transmisión de datos entre aplicaciones de forma flexible, utilizando UDP como protocolo de transporte. Esto lo convierte en una opción ideal para conectar entornos heterogéneos como Python (donde se ejecutan los modelos) y TouchDesigner (donde se generan las visuales). Gracias a OSC, los

diferentes modelos pueden enviar información en tiempo real o de forma puntual según sus necesidades, manteniendo una arquitectura modular y escalable.

5.4.1 Integración – Modelo Predicción de Género musical

En primer lugar, se integrará el modelo de predicción de género musical desarrollado que, como no opera en tiempo real, su funcionamiento se limita a generar una única predicción al inicio de la reproducción de cada canción.

En el caso de este modelo, se ha creado un pequeño script en Python que, una vez generada la predicción de género, empaqueta esta información en un mensaje OSC y lo envía a través del puerto 8001. Durante las pruebas, tanto el modelo como TouchDesigner se han ejecutado en el mismo ordenador, utilizando como dirección IP el localhost (127.0.0.1). Sin embargo, el sistema está diseñado para ser escalable, por lo que, si fuese necesario, podría funcionar también en red, enviando el mensaje a otro ordenador dentro de la misma red local. Para ello, bastaría con cambiar la IP de destino en el script de Python por la del ordenador que ejecuta TouchDesigner, asegurándose de que ambos dispositivos estén conectados a la misma red y que el puerto UDP esté habilitado para la comunicación.

5.4.1.1. Proceso de integración para modelo de género

Como se mencionó anteriormente, se ha desarrollado un script que analiza automáticamente las canciones ubicadas en una carpeta predefinida (configurada por el artista), realiza predicciones de género musical para cada una de ellas y guarda los resultados en un nuevo archivo CSV. Este CSV contiene, para cada canción, las probabilidades de que esta pertenezca a un género u a otro lo que nos da mucha flexibilidad a la hora de crear las visuales:

	Song ID	Predicted Genre	Ambient	Deep House	Techno	Trance	Progressive House
0	Jerro - Demons feat. Sophia Bel (Massane Remix)	Ambient	0.653242290019989	0.3280983567237854	0.0007094230386428	0.0001899467752082	0.017760085735321
1	X2Download.app - Martin Garrix feat. John Martin - Higher Ground (Official Video) (320 kbps)	Progressive House	0.2336292415857315	0.0467351488769054	0.0009238572613887	6.576363375643268e-05	0.7186468243598938
2	Axwell & Shapov - Belong (Axwell & Years Remode) Official Video (320 kbps)	Progressive House	0.1749072074890136	0.0629160329699516	0.0036861796397715	0.0075399852357804	0.7509505152702332

Figura 40. Csv canciones y probabilidades por género

A partir de este archivo CSV, cada vez que el artista cambie de canción se enviará la predicción correspondiente a TouchDesigner. Aunque lo ideal sería contar con un sistema

de identificación automática del audio en tiempo real, esta funcionalidad no forma parte de los objetivos actuales del proyecto. Por ello, durante esta fase y en las pruebas finales, el cambio de canción se realizará manualmente. El mensaje de predicción se envía a Touchdesinger:

```
(.venv) (base) PS C:\Users\administradorlocal\OneDrive - Universidad Pontificia Comillas\TFG\TFG> python Scripts\Communication_genero.py
Enviado a TouchDesigner: /genres Ambient,0.653242290019989,Deep House,0.3280983567237854,Progressive House,0.017760008573532104
Enviado a TouchDesigner: /genres Progressive House,0.7186468243598938,Ambient,0.2336292415857315,Deep House,0.04673514887690544
Enviado a TouchDesigner: /genres Progressive House,0.7509505152702332,Ambient,0.17490720748901367,Deep House,0.06291603296995163
```

Figura 41. Envio Pyhton predicciones - género

Actualmente, el sistema envía las predicciones de tres canciones a TouchDesigner mediante el protocolo OSC. En TouchDesigner se utiliza el componente “oscin DAT” para recibir estos mensajes, y se ha incorporado un script en Python que vuelca las predicciones correspondientes a la primera canción en una “table”, permitiendo su visualización y posterior uso dentro del entorno visual.

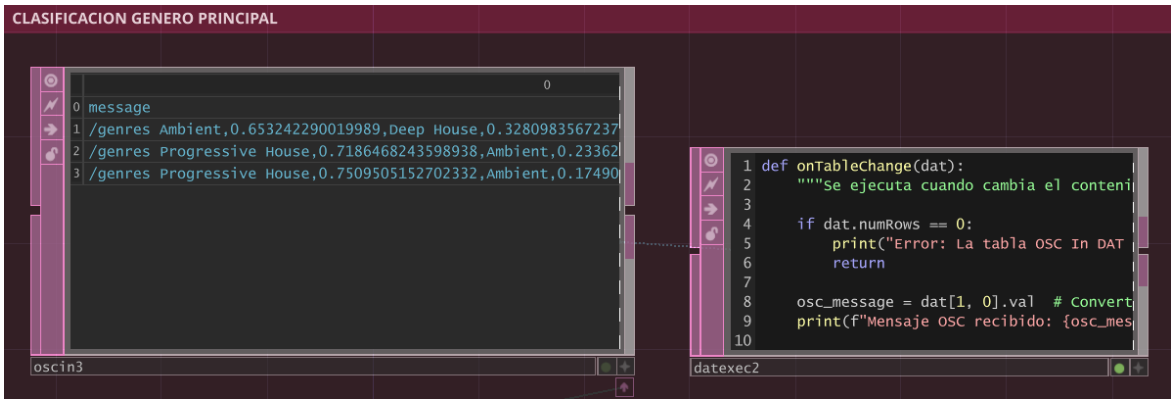

Figura 42. Recepción Touchdesigner predicciones - género

En la tabla generada en TouchDesigner se muestra, por fila, cada género asociado a la canción y su probabilidad correspondiente.

	0	1
0	Genre	Probability
1	Ambient	0.653242290019989
2	Deep House	0.3280983567237854
3	Progressive House	0.017760008573532104

Figura 43. Tabla de probabilidades en Touchdesigner – género

Diseño de visuales en función de las predicciones

La lógica propuesta para generar las visuales a partir de las predicciones es la siguiente:

- **Género principal (el de mayor probabilidad):** define el sistema de mapeo y análisis de audio que se aplicará a las visuales. Por ejemplo, si el género principal es Ambient, se utilizará una configuración de audio adaptada específicamente a ese estilo, con visuales suaves y reactivas a las frecuencias predominantes del género.

- **Género secundario (segunda predicción más probable):** determina la base visual sobre la que se construirá la escena. Para cada género principal existen dos posibles bases visuales; se seleccionará una de ellas en función del género secundario. Además, según las probabilidades asociadas al primer y segundo género, se podrán aplicar variaciones adicionales como la elección de diferentes "seeds" o trazados de líneas dinámicos.

- **Tercer género (tercera predicción más probable):** afecta al brillo general de la visual. Por ejemplo, si el tercer género es Progressive House, se configurará un brillo de intensidad media-alta, adaptado a las características visuales típicas de este estilo.

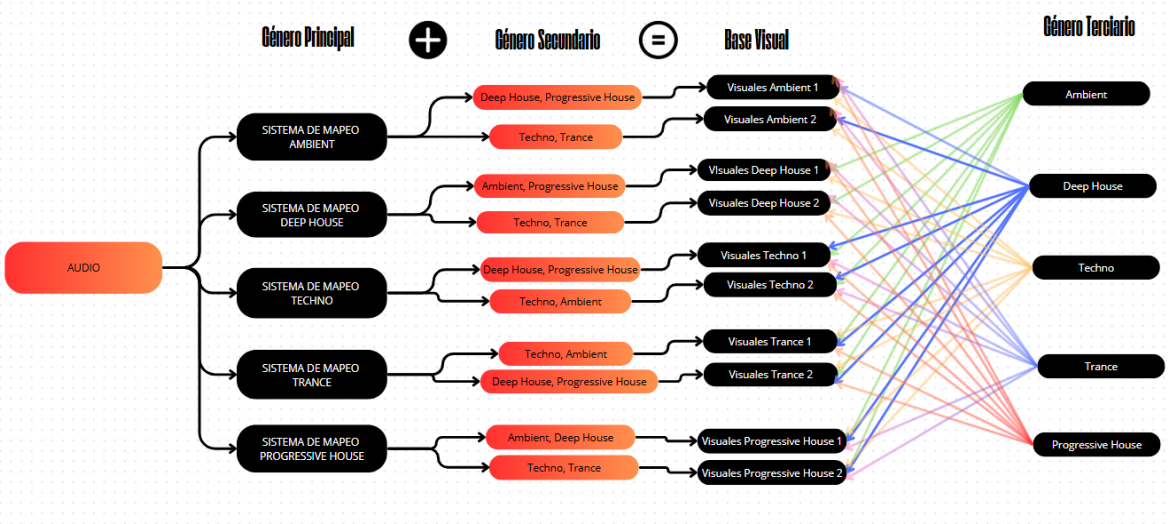

Figura 44. Esquema lógico generación de visuales en función de géneros

5.4.1.2. Resultados visuales modelo de género

Ejemplo 1:

Género Principal: Deep House, Género Secundario: Techno, Género terciario: Ambient.

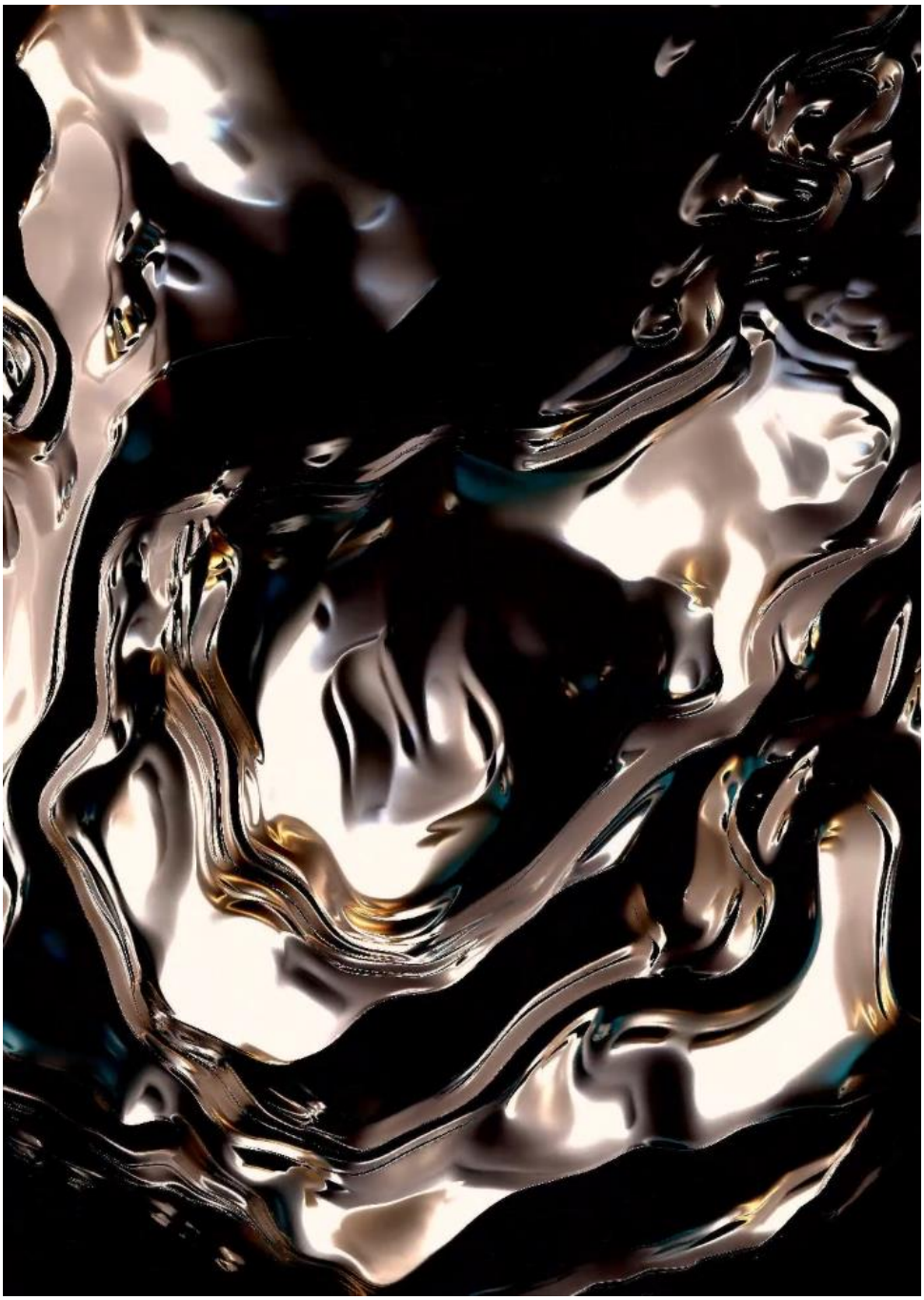

Figura 45: Visual ejemplo1 - género

Ejemplo 2:

Género Principal: Deep House, Género Secundario: Ambient, Género Terciario: Trance.

Figura 46: Visual ejemplo 2 - género

Ejemplo 3:

Género Principal: Progressive House, Género Secundario: Deep House, Género Terciario: Techno.

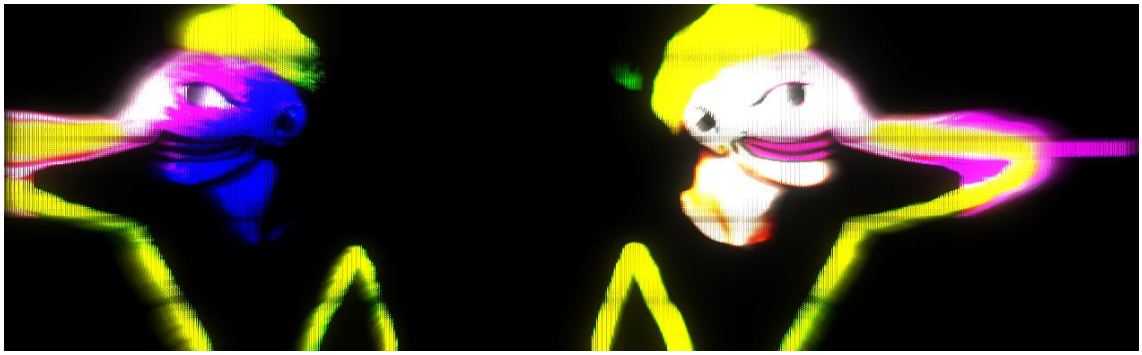

Figura 47: Visual ejemplo 3 - género

Ejemplo 4:

Género Principal: Techno, Género Secundario: Trance, Género Terciario: Deep House.

Figura 48: Visual ejemplo 4 - género

Ejemplo 5 (Visuales canciones muy parecidas):

- Visual 1: Género Principal: Ambient, Género Secundario: Deep House, Género Terciario: Deep House.
- Visual 2: Género Principal: Ambient, Género Secundario: Progressive House, Género Terciario: Deep House.

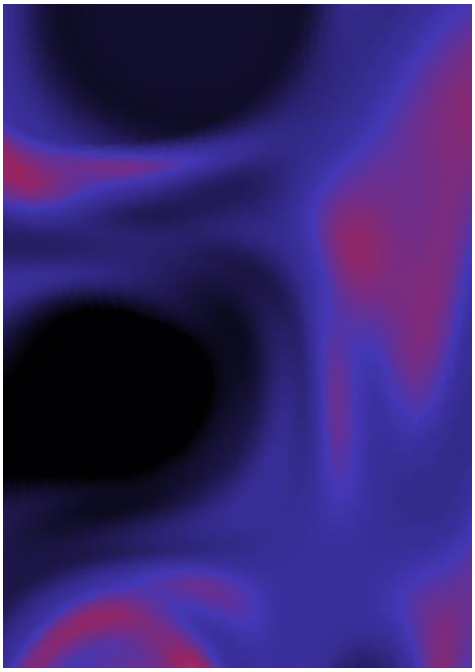

Figura 49: Visual 1 ejemplo 4 -género

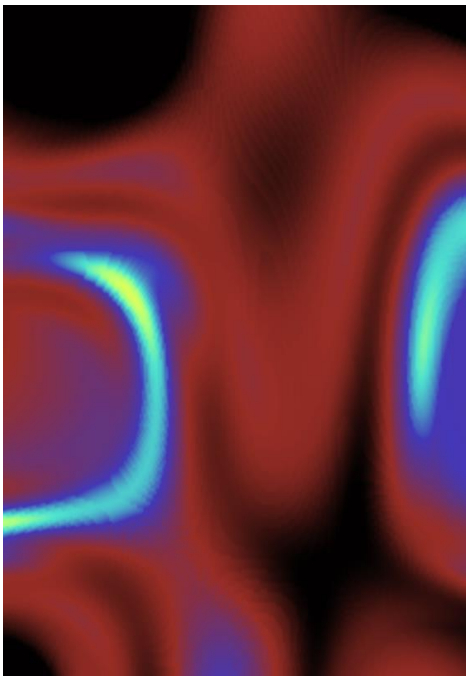

Figura 50: Visual 2 ejemplo 4 - género

5.4.2 Integración – Modelo de Predicción de Emociones:

Para el modelo de predicción de emociones el protocolo de comunicación utilizado también es el OSC. En este caso las predicciones se realizan durante la reproducción de la canción en ventanas de 5 segundos, por lo que ahora se enviara un mensaje OSC desde Python a Touchdesigner cada 5 - 6 segundos debido al tiempo de procesamiento en la predicción del modelo.

Al igual que en el anterior se ha configurado un script en Python que hace la predicción y se la comunica a Touchdesigner:

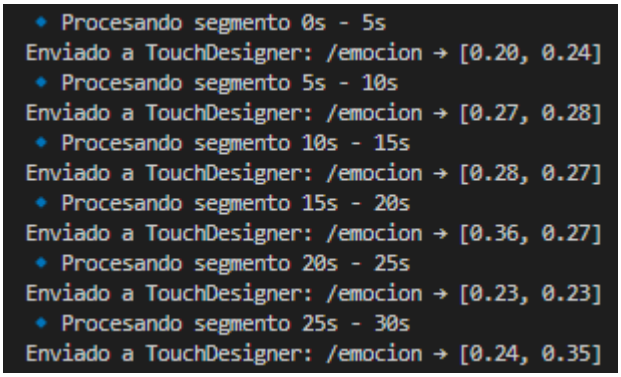

Figura 51. Envió Python predicciones - emociones

En este caso, al ser dos valores para una misma canción que van a ir variando los recibimos en Touchdesigner a través de un "oscin CHOP".

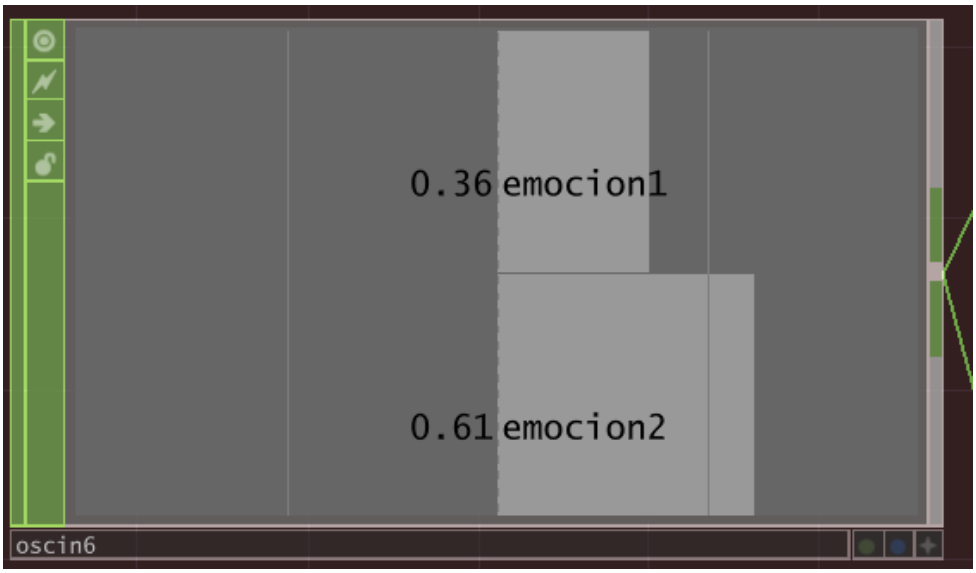

Figura 52. Recepción Touchdesigner predicciones - emociones

Una vez recibidos los valores, los separamos en canales individuales. En este caso se ha decidido hacer dos mapeos a la visual:

Brillo: Por una parte, cogemos el valor de arousal y lo mapeamos al brillo de las visuales. La lógica es simple, cuando el arousal sea alto la visual se volverá más brillante. Para un mejor resultado aplicamos un "Lag CHOP". Este componente permite suavizar las transiciones entre valores, evitando cambios bruscos. En lugar de actualizarse de forma

inmediata, los valores se ajustan de manera progresiva, generando transiciones más fluidas y naturales en la visual.

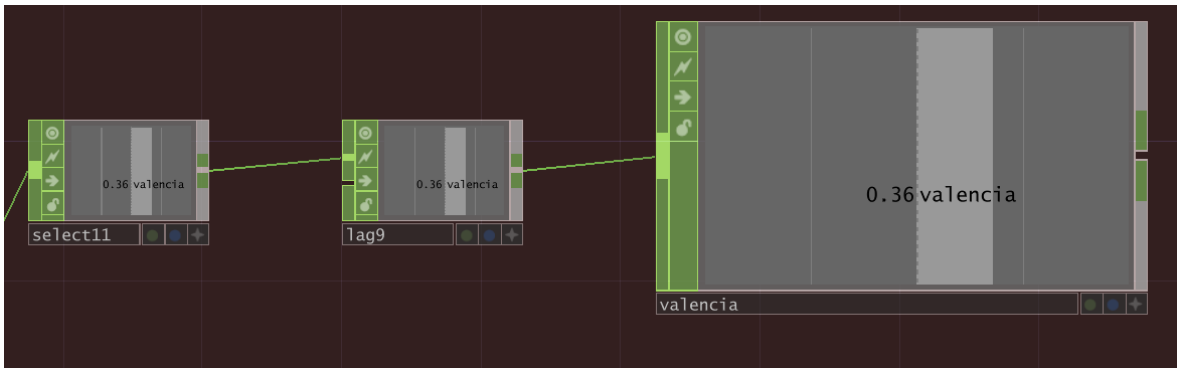

Figura 53. "Lag Chop" suavizado de transición

Color: La representación visual del color se adapta dinámicamente en función de las emociones percibidas en la música. Para ello, se realiza una clasificación emocional a partir de la suma de los valores de valencia y arousal mediante un nodo "Math". Este valor combinado determina qué gama cromática se utiliza en cada momento.

Se han diseñado distintas rampas de color con variaciones sutiles entre ellas, y cada una está asociada a un estado emocional concreto. Por ejemplo, si una canción transmite alegría, se asigna automáticamente una rampa cálida y luminosa; si en cambio la emoción dominante es la calma, se activa una rampa más suave y serena.

Este sistema permite que las transiciones de color respondan de forma automática y fluida a los cambios emocionales de la música, aportando coherencia y riqueza visual al contenido.

5.4.2.1. Resultados visuales modelo de emociones

Brillo:

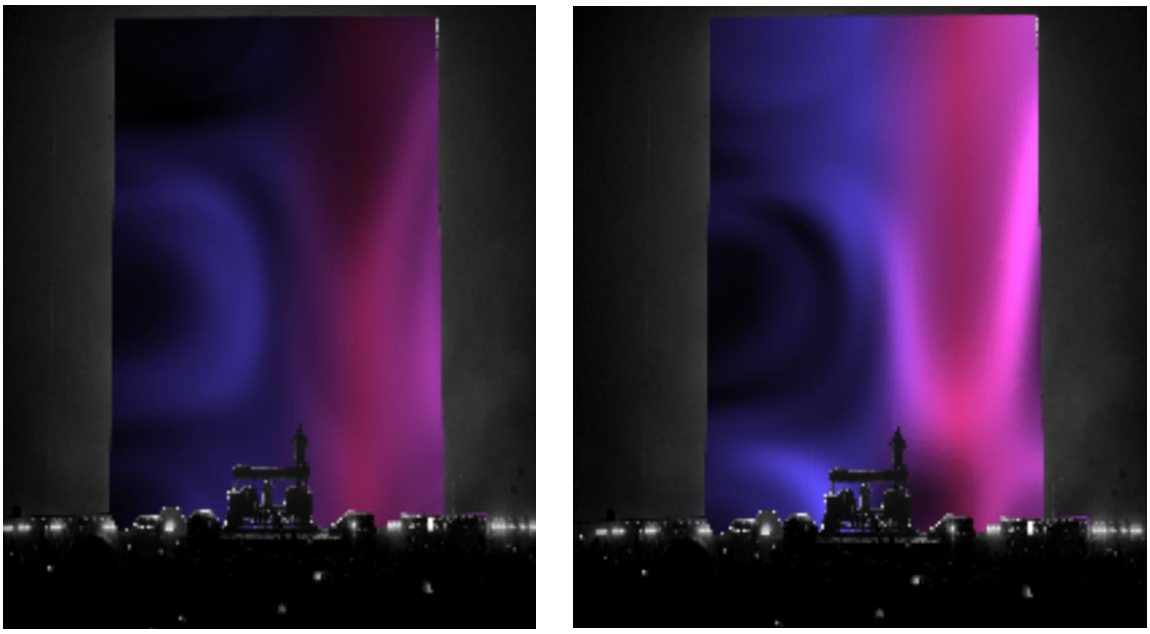

Figura 54. Visual brillo – emociones

Colores:

Como se he explicado se crean paletas manteniendo la misma gama de colores pero que puedan representar diferentes estados de ánimo:

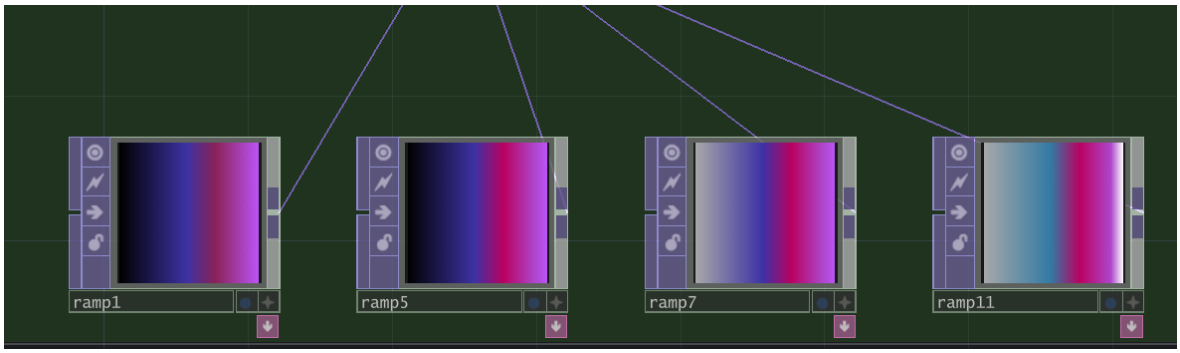

Figura 55. Paletas de colores – visuales

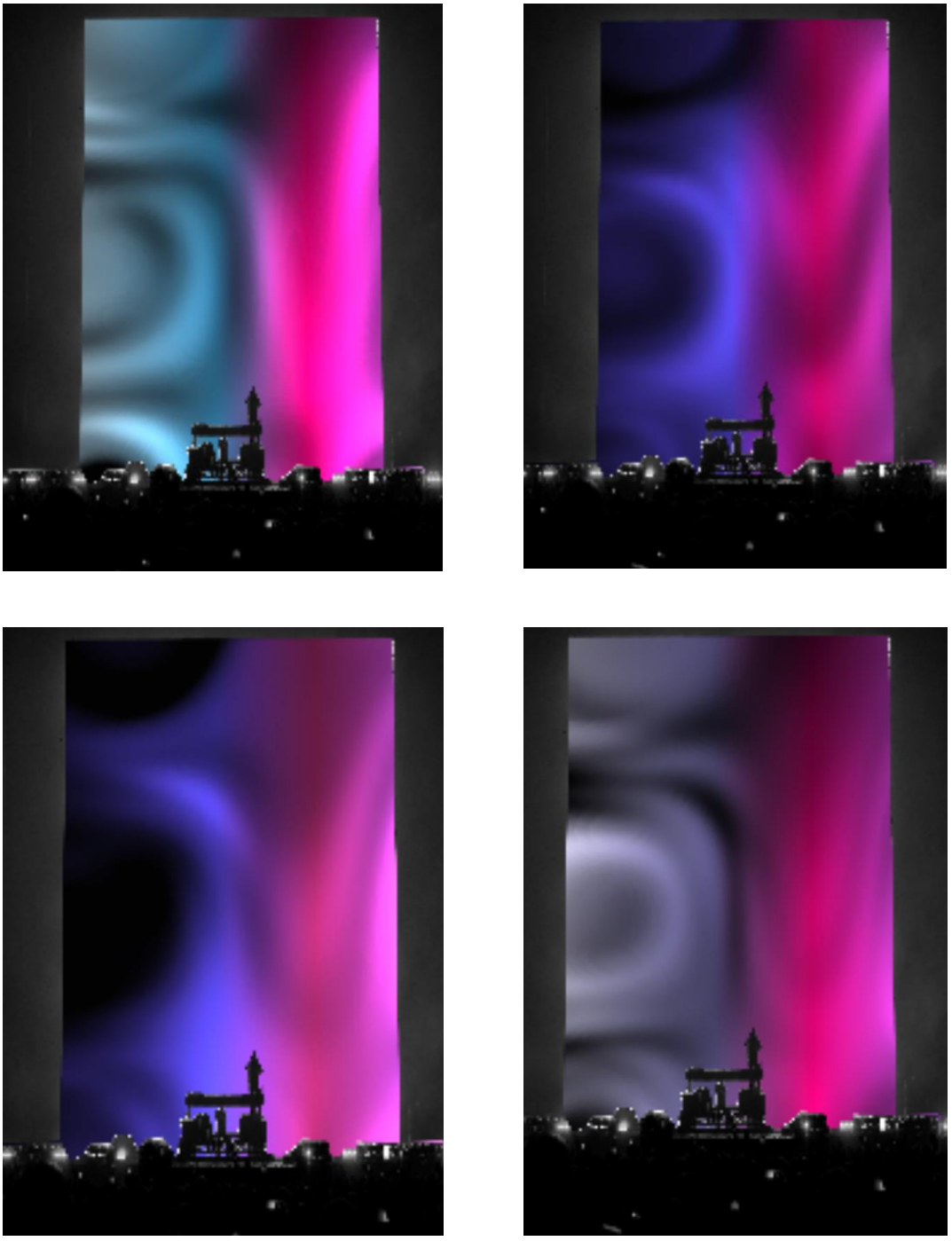

Figura 56. Cambio de color en visuales – emociones

5.4.3 Integración – Modelo de Predicción de Secciones

Este último modelo es el más fácil de integrar en TouchDesigner, ya que solo hace falta un pequeño script en Python que envíe un 1 cuando detecta un pre-drop en la canción. Con ese valor, las visuales pueden anticiparse al drop y activar distintos efectos justo en ese momento.

La conexión con TouchDesigner también es simple. Se usará un nodo llamado "oscillator in", que se encarga de recibir ese 1. A partir de ahí, se pueden encender diferentes elementos visuales para que reaccionen a la predicción.

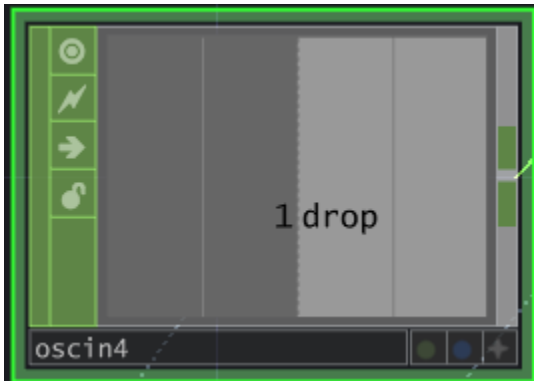

Figura 57. Recepción Touchdesigner predicción - sección

Dependiendo del tipo de visual, la reacción puede variar. En este proyecto, para probar cómo se comporta el sistema, se han creado dos formas distintas de respuesta:

- **Vibración:** Cuando llega la predicción del drop, la visual empieza a vibrar. Esto se consigue conectando un nodo "constant", que es una constante (en este caso 1) al "oscillator" y multiplicando su valor por una señal que varía entre 0 y 1. Antes de que llegue la señal de pre-drop, el resultado de la multiplicación es 0, por lo que no hay movimiento. Pero en cuanto entra el 1, esa multiplicación activa la vibración usando la señal variable como intensidad.

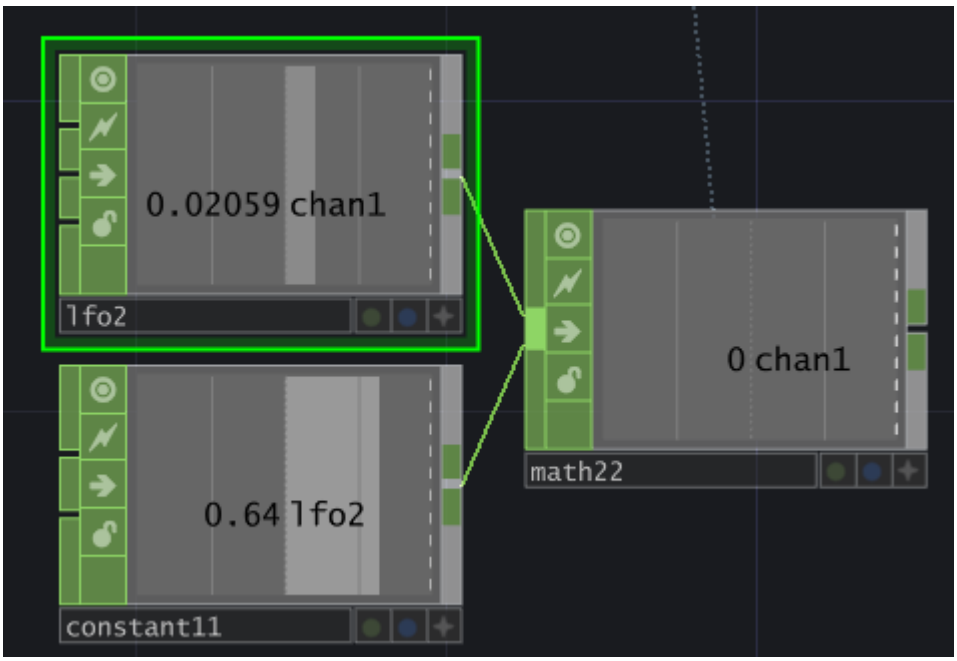

Figura 58. Sistema lógico vibración – sección

El output de este sistema se mapea al parámetro "Transalate" de la visual lo que hace que vibre rápidamente:

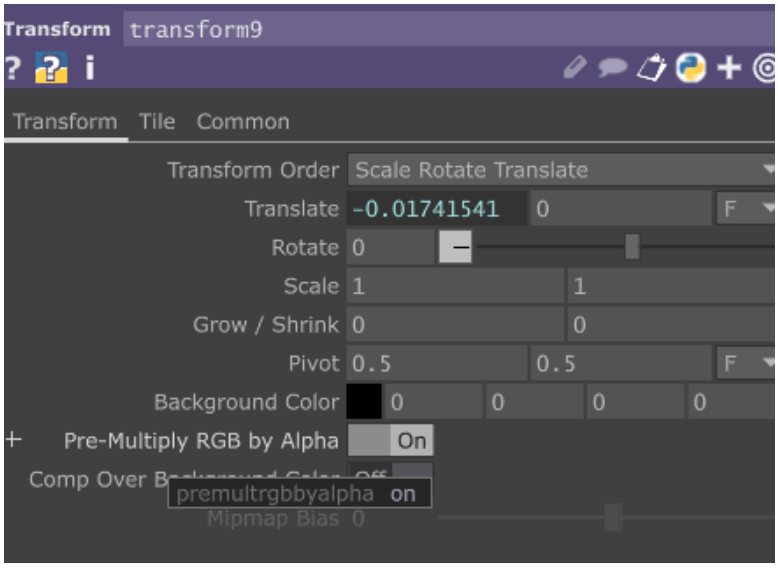

Figura 59. Cambio parámetro "Translate" – sección

Ejemplo 1:

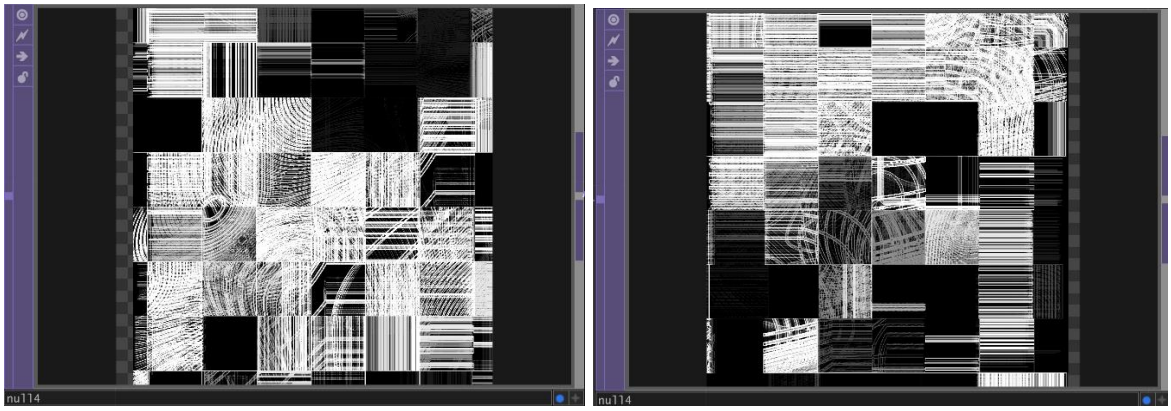

Figura 60. Ejemplo 1 vibración visual - sección

- **Desvanecimiento:** Para este funcionamiento la integración es muy similar. En este caso se suma el 1 o el 0 recibida por una constante. Mientras no se recibe un 1 se mantiene el 0 y la visual no cambia. Sin embargo, cuando se recibe el 1 se activa el desvanecimiento de la visual.

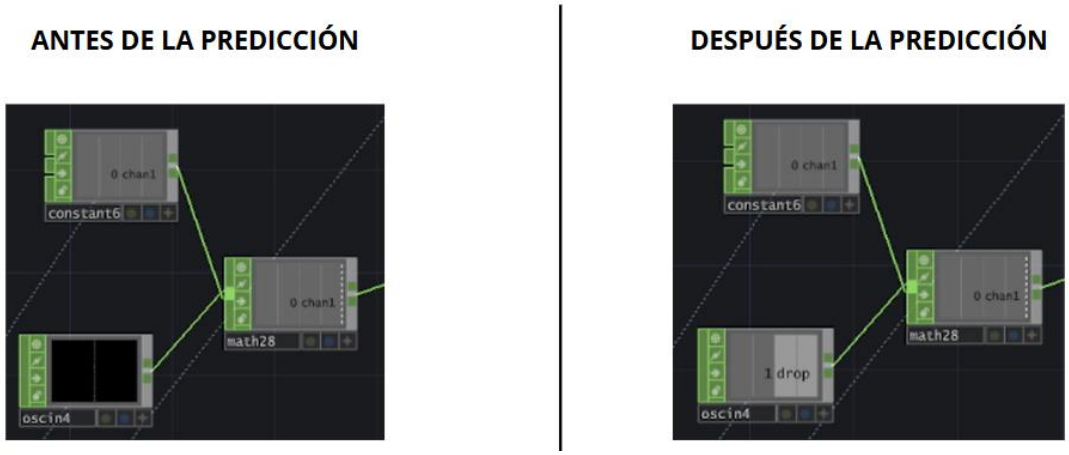

Figura 61. Lógica recepción de pre-drop – sección

5.4.3.1. Resultados visuales modelo de secciones

Ejemplo 1:

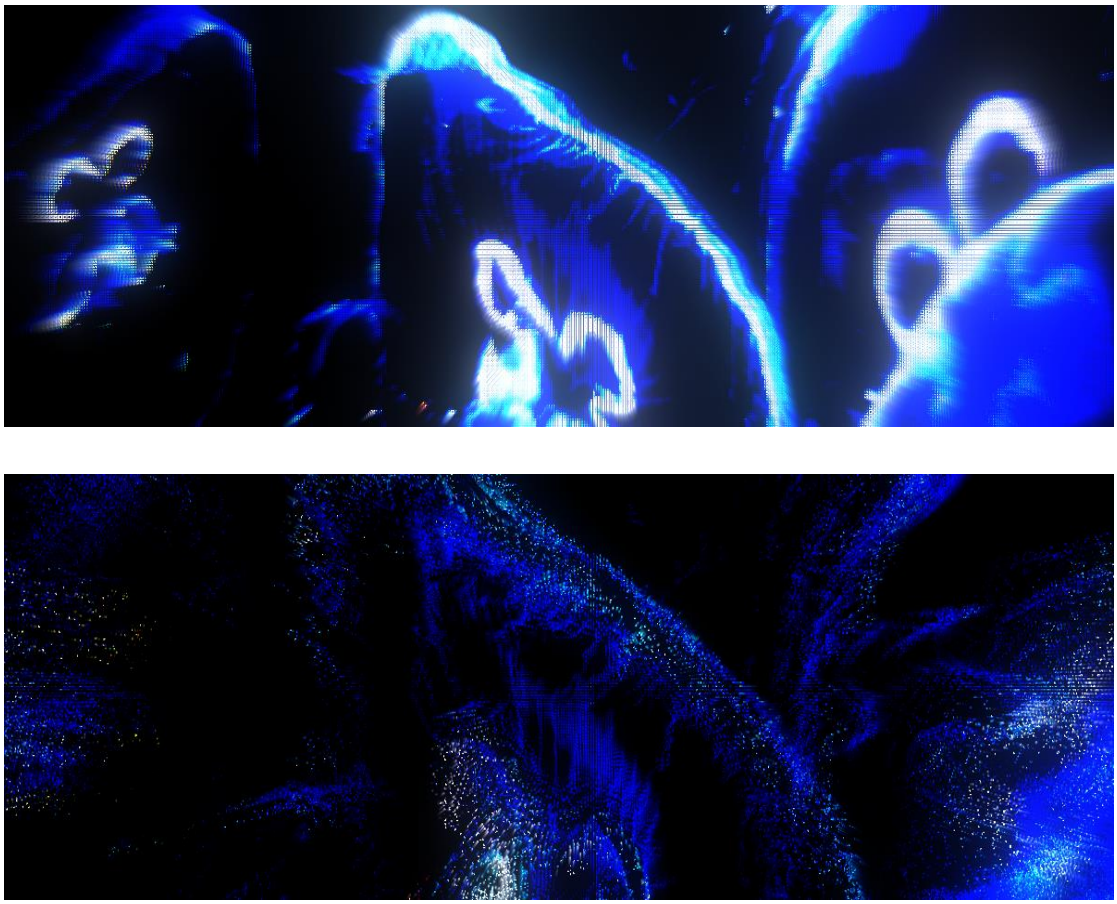

Figura 62: Desvanecimiento ejemplo1 - sección

Ejemplo 2:

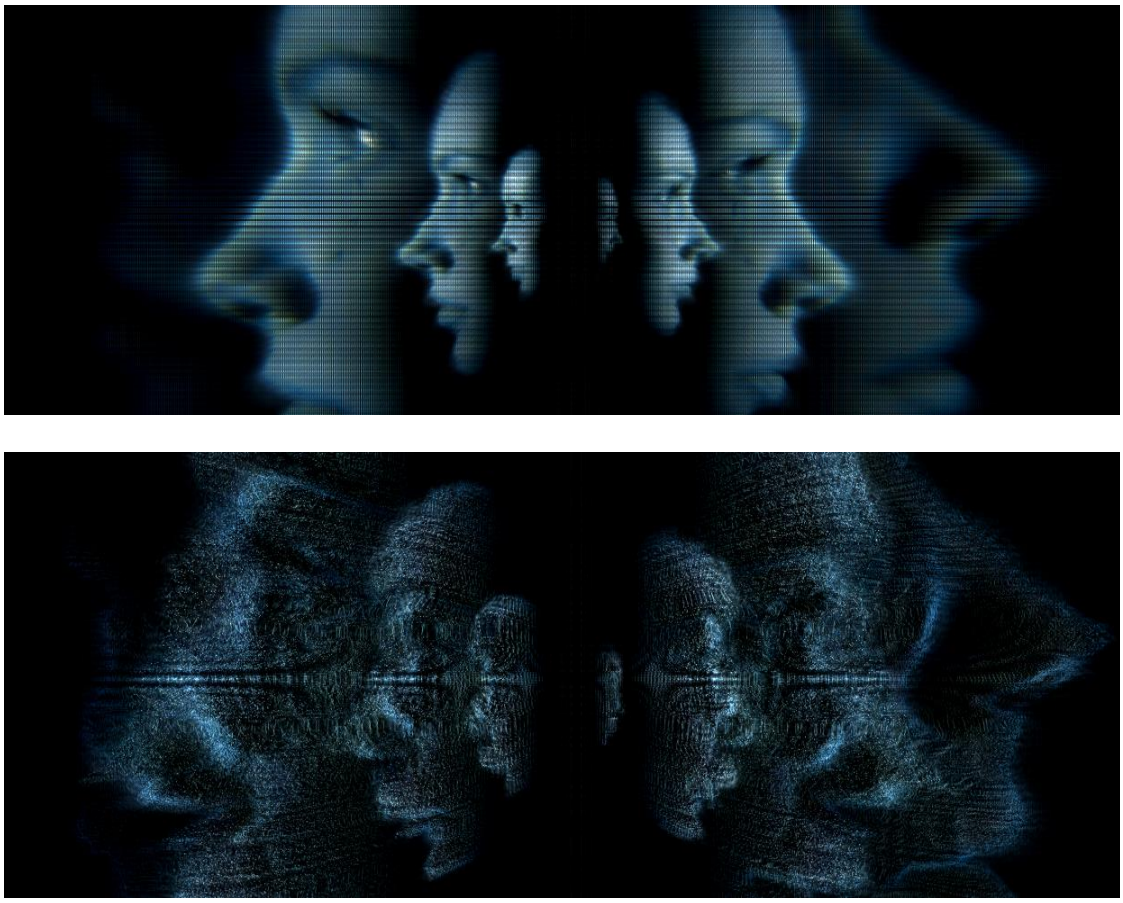

Figura 63: Desvanecimiento ejemplo 2 – sección

Ejemplo 3: (Imagen de Dillon Francis)

Figura 64: Desvanecimiento ejemplo 3 - sección

Capítulo 6. ANÁLISIS DE RESULTADOS

En el siguiente capítulo se hablará de los resultados más relevantes del trabajo final y de cada uno de los modelos.

6.1. ANÁLISIS RESULTADOS EN MODELO DE GÉNERO

Los resultados obtenidos con el modelo de clasificación de género musical pueden considerarse satisfactorios, tanto a nivel técnico como en su integración con el sistema de visuales. La arquitectura final, basada en una combinación de redes CNN y GRU, ha ofrecido un rendimiento notable dadas las condiciones del problema: una clasificación fragmentada por tramos de 15 segundos, lo cual introduce una dificultad añadida al análisis. A pesar de ello, las métricas obtenidas —con una accuracy superior al 64% en validación— confirman que el modelo ha aprendido patrones significativos en los datos, y que las predicciones tienen un valor representativo fiable, especialmente cuando se agregan a nivel de canción.

Además del rendimiento en validación, el análisis con un conjunto de canciones más ambiguas y estilísticamente híbridas ha permitido comprobar la utilidad real del enfoque probabilístico planteado. El diseño de modelo que asigna probabilidades frente a una etiqueta fija se ha revelado especialmente eficaz al integrarse con el motor de visuales, permitiendo generar escenas más ricas y adaptativas en función del perfil sonoro de cada pista. Incluso en los casos en que el modelo comete errores, estos suelen producirse entre géneros similares, lo que sugiere que las confusiones son razonables y coherentes con la naturaleza subjetiva y mixta de la música electrónica.

En lo que respecta a la integración técnica, los resultados también han sido positivos. El sistema diseñado para gestionar la comunicación entre el modelo y TouchDesigner ha demostrado ser funcional y eficiente, sin introducir latencias perceptibles en el flujo de trabajo. La lógica desarrollada para interpretar las probabilidades —asignando roles distintos

a los tres géneros con mayor peso— se ha mostrado versátil y práctica, permitiendo una manipulación visual rica a partir de una entrada relativamente simple.

No obstante, cabe señalar que, si bien el sistema actual funciona correctamente, todavía requiere intervención manual para preparar los archivos y adaptar las plantillas visuales a cada género. Una mayor automatización en esta parte del pipeline sería deseable, ya que permitiría escalar el sistema sin necesidad de realizar ajustes manuales cada vez que se incorpora nueva música. Asimismo, aunque la lógica ha sido pensada para poder combinar múltiples géneros de forma flexible, no se ha llegado a probar con todas las combinaciones posibles debido a limitaciones de tiempo.

En conjunto, el modelo de género no solo cumple con los objetivos funcionales del proyecto, sino que también sienta unas bases sólidas para futuras mejoras. Su enfoque flexible, su razonable precisión y su integración fluida con el entorno visual lo convierten en una herramienta útil y con gran potencial para espectáculos adaptativos en tiempo real.

6.2. ANÁLISIS DE RESULTADOS EN MODELO DE EMOCIONES

El sistema de predicción emocional desarrollado ha demostrado ser una herramienta funcional y efectiva dentro del ecosistema general de visuales reactivas. A nivel cuantitativo, el rendimiento del modelo Random Forest en la predicción de valencia y arousal ha sido sólido, con errores bajos y coeficientes de determinación que evidencian una buena capacidad explicativa. Esto es especialmente relevante si se considera que el etiquetado emocional en música conlleva una alta carga de subjetividad, ambigüedad y variabilidad temporal. En ese contexto, haber alcanzado estos niveles de precisión indica que el modelo ha logrado capturar patrones emocionales consistentes en los datos. A continuación, se presentan ejemplos aplicados a diferentes canciones que permiten observar cómo el modelo va generando sus predicciones de manera continua conforme avanza el audio reconociendo diferentes emociones:

High on Life- Martin Garrix (Progressive House):

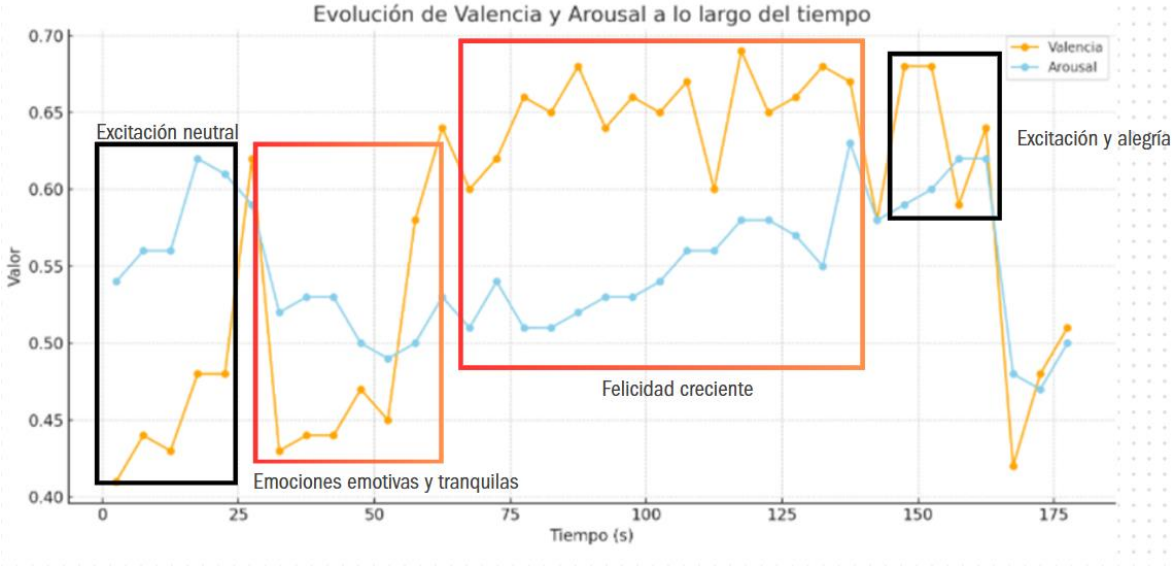

Figura 65. Evolución de Valencia y Arousal de High on Life

La canción High on Life de Martin Garrix podría encuadrar dentro del género progressive house. Se caracteriza por una estructura emocionalmente ascendente: comienza con una atmósfera suave, nostálgica y emotiva, dominada por vocales melódicas y acordes cálidos, y va construyendo gradualmente hacia una sección de alta energía y euforia.

El modelo de predicción captura con bastante precisión esta evolución emocional. Durante los primeros compases, detecta emociones asociadas a la melancolía y nostalgia, reflejando el tono íntimo de la introducción y los versos. Conforme se acerca el primer drop —alrededor de los 110-125 segundos—, las predicciones muestran un aumento progresivo en los niveles de felicidad e intensidad.

El drop de High on Life es una explosión de optimismo, liberación emocional y energía positiva, características que nuestro modelo también identifica bastante bien con una alta valencia y una media-alta arousal.

I´m losing it- Fisher (Techno):

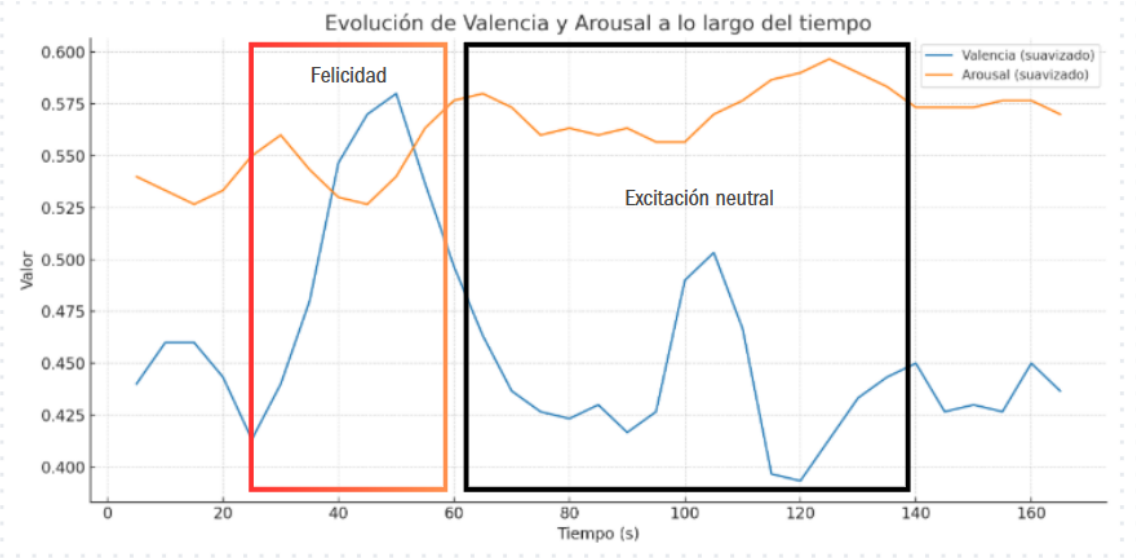

Figura 66. Evolución de Valencia y Arousal de I´m losing it

La canción I'm Losing It de Fisher se enmarca dentro del género Tech House, caracterizándose por un ritmo hipnótico, repetitivo y un enfoque minimalista en su construcción sonora. A diferencia de High on Life, esta pista no busca una evolución emocional marcada, sino mantener una tensión constante en la pista de baile.

El modelo predice correctamente esta dinámica. Los valores de valencia (asociados a la positividad o "felicidad" de la emoción) se mantienen en niveles bajos o moderados, reflejando el tono más oscuro y crudo de la producción. Por otro lado, el arousal (nivel de activación o energía) es medio-alto y muy constante a lo largo de toda la canción, sin grandes picos ni caídas.

Esto se corresponde con la estructura musical de I'm Losing It, donde predomina un groove repetitivo que sostiene la energía de manera estable, sin recurrir a cambios dramáticos de intensidad ni momentos de explosión emocional. Así, el modelo captura fielmente el objetivo principal de este tipo de track: generar un estado de trance rítmico y mantener al oyente atrapado en una atmósfera constante y envolvente.

Begin Again – Ben Bohmmer (Ambient):

Figura 67. Evolución de Valencia y Arousal de Begin Again

En Begin Again, una canción del género Ambient, el modelo predice una evolución emocional coherente con la estructura musical de la canción. Durante la primera sección, con una base ambiental y suave, se observan valores bajos de valencia y arousal, transmitiendo una sensación de calma. A partir de los 30 segundos, el arousal aumenta y se mantiene estable en niveles medios-altos, acompañando la entrada de ritmos más marcados y una mayor intensidad musical, que transmite una sensación animada y en construcción.

Entre los 90 y 140 segundos, tanto la valencia como el arousal alcanzan sus valores más altos, reflejando una fase de alegría y optimismo, coincidiendo con el clímax emocional de la canción. Finalmente, en los últimos segundos, ambos valores descienden bruscamente, volviendo a una atmósfera de calma que cierra el viaje emocional.

A pesar de que los resultados obtenidos son positivos, el modelo presenta ciertas limitaciones. En canciones con estructuras más complejas o menos convencionales, su rendimiento tiende a ser menos consistente. Tal como se ha mencionado anteriormente, el número reducido de muestras de entrenamiento dificulta la capacidad de generalización del

modelo, lo que repercute en su eficacia ante situaciones menos representadas durante el entrenamiento.

Más allá de las métricas, la integración del modelo con el entorno visual ofrece resultados perceptibles y coherentes en la práctica. La elección de una ventana temporal de 5 segundos para la actualización de valores, junto con el uso del protocolo OSC, permite una comunicación fluida entre el sistema de inferencia y TouchDesigner, consiguiendo una experiencia visual dinámica pero no caótica. Las decisiones de diseño en la traducción de emociones a visuales —como el uso del arousal para controlar el brillo y de la suma de arousal y valencia para determinar el color— son simples pero efectivas, logrando que la visualización reaccione de forma intuitiva al estado emocional transmitido por la música.

Sin embargo, también se identifican ciertos límites en la implementación actual. Aunque el sistema de rampas cromáticas ofrece una paleta adaptable a distintos estados emocionales, su aplicación sigue siendo manual y dependiente del diseño previo de cada escena. Esto introduce una rigidez que podría limitar la escalabilidad o la adaptabilidad del sistema en contextos más complejos o automatizados. Además, si bien el uso de herramientas como el "Lag CHOP" suaviza la transición de valores y contribuye a una mejor estética, el pipeline general sigue siendo relativamente lineal, sin una capa de interpretación más abstracta o semántica de las emociones.

En resumen, el módulo emocional cumple su propósito de enriquecer la experiencia visual mediante información afectiva extraída del audio, aportando variedad y coherencia a las escenas generadas. Su estructura actual lo convierte en un buen punto de partida funcional, pero también abre la puerta a futuras mejoras en términos de autonomía, sofisticación interpretativa y adaptabilidad visual. La posibilidad de explorar técnicas más avanzadas, como redes recurrentes o modelos de atención, o integrar directamente las emociones en la lógica generativa de visuales, representa una oportunidad clara para evolucionar el sistema hacia propuestas más inmersivas y expresivas.

6.3. ANÁLISIS RESULTADOS EN MODELO DE SECCIONES

Los resultados obtenidos con el modelo de detección de pre-drop, si bien no destacan por su precisión global, han demostrado ser adecuados para los fines prácticos del proyecto. Se trata de un modelo de arquitectura sencilla, entrenado para emitir una señal binaria cuando detecta la inminencia de un drop en la canción. Esta simplicidad ha resultado ser una ventaja en términos de integración y eficiencia, ya que permite una respuesta rápida y directa sin requerir un procesamiento complejo o intensivo.

Desde el punto de vista técnico, el principal obstáculo ha sido la limitada cantidad y diversidad de datos disponibles para el entrenamiento. Esto ha afectado directamente a la capacidad del modelo para generalizar, provocando errores en canciones que se alejan de las estructuras más convencionales. Aun así, el comportamiento observado en pruebas reales muestra que, cuando se producen predicciones correctas, estas son coherentes y oportunas, lo que refuerza su valor como herramienta creativa. Dado que no se trata de un sistema que deba clasificar constantemente, sino de actuar en momentos clave, incluso una tasa moderada de aciertos puede tener un gran impacto visual.

En cuanto a la integración con TouchDesigner, los resultados han sido especialmente positivos. y han desarrollado dos respuestas principales: una vibración, que introduce una sensación de tensión justo antes del drop, y un desvanecimiento progresivo, que aporta una transición suave entre escenas visuales.

Ambos efectos han sido diseñados para amplificar la experiencia sonora desde el plano visual, y en las pruebas realizadas han cumplido este objetivo de forma eficaz. La vibración aporta dinamismo e inmediatez, mientras que el desvanecimiento genera una atmósfera más envolvente y narrativa. Pese a que las activaciones no siempre coinciden con momentos óptimos, el resultado visual cuando lo hacen es muy convincente y potencia claramente la experiencia del espectador.

Cabe destacar que la principal limitación del sistema no radica en la arquitectura del modelo, sino en la escasez de datos y la falta de diversidad en el conjunto de entrenamiento, un aspecto que puede solucionarse fácilmente en el futuro. Esto convierte al modelo en una base sólida sobre la que seguir iterando, mejorando progresivamente la sensibilidad sin necesidad de rehacer su diseño ni su integración.

En conjunto, el sistema de detección de pre-drop cumple su propósito funcional dentro del proyecto: permitir que las visuales reaccionen en tiempo real a la música de forma contextual y expresiva. Aunque queda margen de mejora, especialmente en la precisión del modelo, la experiencia lograda con las predicciones correctas es potente y demuestra el potencial del enfoque adoptado. Se trata, por tanto, de una solución viable y con recorrido, que sienta unas bases claras para el desarrollo de visuales más inteligentes y sincronizadas en espectáculos en directo.

6.4. ANÁLISIS RESULTADOS EN MODELO DE AUDIO

El modelo de análisis de audio ha demostrado un desempeño eficiente, reaccionando con rapidez a las variaciones musicales. Los filtros de frecuencia implementados capturan con precisión los diferentes momentos de la canción, así como los cambios en la energía y elementos clave como los bombos, permitiendo que las visuales respondan de forma dinámica y sincronizada en tiempo real.

Los modelos desarrollados han cumplido de forma razonable con los objetivos propuestos, ofreciendo un rendimiento aceptable dadas las limitaciones del conjunto de datos y del tiempo disponible. Aunque algunas métricas de precisión no alcanzan niveles óptimos, los modelos han demostrado ser funcionales en contextos reales, especialmente en su integración con sistemas visuales.

Capítulo 7. CONCLUSIONES Y TRABAJOS FUTUROS

Este Trabajo de Fin de Grado ha demostrado que la integración de inteligencia artificial en el ámbito de los espectáculos visuales es una posibilidad real y con gran potencial de desarrollo. A pesar de las limitaciones propias del proyecto —como el tiempo limitado para su realización, los recursos técnicos disponibles y la necesidad de coordinar distintas áreas del conocimiento— se ha logrado construir un sistema funcional que enlaza modelos de Deep Learning con entornos de generación visual en tiempo real.

Uno de los éxitos más importantes ha sido confirmar la viabilidad de usar redes neuronales entrenadas con fragmentos de audio para crear visuales reactivas que respondan de manera contextual durante una actuación en vivo. Aunque el enfoque sigue siendo experimental, el empleo de técnicas de aprendizaje profundo aplicadas al audio ha producido resultados alentadores, sobre todo al detectar elementos clave como patrones rítmicos, cambios de intensidad o drops en la música.

El proyecto también ha aportado novedades importantes. Se ha implementado un flujo de trabajo integral que conecta el análisis automático de audio con la generación visual en TouchDesigner, garantizando una comunicación eficiente entre ambos sistemas. Se han probado diferentes arquitecturas híbridas, como la combinación de CNN con GRU y Random Forest, que han mostrado buena capacidad para funcionar con datos limitados, sentando una base firme para futuras optimizaciones. Además, se ha desarrollado una lógica versátil que traduce las salidas probabilísticas del modelo en comandos visuales dinámicos, abriendo la puerta a aplicaciones más creativas y flexibles en entornos escénicos.

7.1. OBJETIVOS ESTABLECIDOS

En cuanto a los objetivos establecidos, la mayoría han sido alcanzados de manera efectiva. Se ha desarrollado un software modular capaz de analizar fragmentos de canciones, predecir eventos relevantes y traducir esas predicciones en respuestas visuales controladas mediante TouchDesigner. Aunque el sistema aún no se encuentra en una fase plenamente operativa para su uso profesional, su arquitectura y diseño permiten prever un desarrollo funcional con ajustes adicionales, mejoras en el preprocesamiento y una ampliación del conjunto de datos.

Además, el carácter original del proyecto ha exigido un enfoque basado en prueba y error, exploración activa y toma de decisiones creativas en cada etapa del desarrollo. Lejos de seguir una plantilla existente, se ha trabajado desde una idea innovadora que combina música, inteligencia artificial y arte visual, lo que ha implicado una fuerte componente investigadora. Esta singularidad no solo ha sido un desafío técnico, sino también una fuente de motivación para abordar el problema desde una perspectiva interdisciplinar.

7.2. TRABAJOS FUTUROS

En conjunto, el trabajo ha servido como un punto de partida sólido para un campo con un enorme potencial: la generación de experiencias audiovisuales adaptativas mediante inteligencia artificial. Lo desarrollado representa una primera aproximación funcional a lo que podría convertirse en un software profesional tradicional para espectáculos en vivo. Las herramientas y enfoques empleados son escalables, y con un mayor volumen de datos, una lógica visual más sofisticada y una configuración más robusta, el sistema tiene margen de evolución hacia un producto real aplicable en entornos profesionales.

Las principales líneas de mejora de cara al futuro se centran en tres ejes. En primer lugar, la ampliación del dataset y el refinamiento de los modelos de Deep Learning permitirían lograr predicciones más precisas y robustas en tiempo real. En segundo lugar, el sistema de configuración debería automatizarse completamente para eliminar cualquier necesidad de intervención manual, permitiendo su uso inmediato en directo por parte de artistas sin

conocimientos técnicos. Y, en tercer lugar, la generación visual, si bien funcional, aún puede enriquecerse incorporando mayor variedad estética, más parámetros dinámicos y una lógica creativa más compleja que aproveche en mayor profundidad las predicciones del modelo.

Lo que hoy es un prototipo funcional, mañana puede ser una herramienta creativa disruptiva para artistas sin grandes recursos técnicos, ampliando las posibilidades expresivas del directo y democratizando el acceso a visuales inteligentes. Las bases ya están sentadas; lo que queda por delante es un trabajo de expansión, refinamiento y diseño para convertir esta visión en una realidad plenamente operativa.

Capítulo 8. BIBLIOGRAFÍA

Bittner et al. Bittner, R. et al. "Two-dimensional valence-arousal space for emotion representation." Investigación clave que establece el modelo valencia-arousal para clasificar emociones, fundamental en la interpretación emocional en audio y visuales. https://www.researchgate.net/figure/Two-dimensional-valence-arousal-space_fig1_304124018

Doshi, Ketan.
Doshi, K. "Audio Deep Learning – Sound Classification." Guía introductoria que explica técnicas de deep learning para la clasificación de sonidos basada en características acústicas. https://medium.com/data-science/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5

Oluyale, David.
Oluyale, D. "Audio Classification Using Deep Learning." Tutorial práctico sobre la implementación de modelos de deep learning para clasificación de audio utilizando TensorFlow. https://medium.com/@oluyaled/audio-classification-using-deep-learning

Analytics Vidhya. "Music Genres Classification Using Deep Learning Techniques." Explicación de métodos avanzados para clasificar géneros musicales con redes neuronales profundas. https://www.analyticsvidhya.com/blog/2021/06/music-genres-classification-using-deep-learning-techniques/

DataCamp. "Introduction to Convolutional Neural Networks (CNNs)." Tutorial básico sobre CNNs, esenciales para el análisis de espectrogramas en aplicaciones de audio y visuales. https://www.datacamp.com/es/tutorial/introduction-to-convolutional-neural-networks-cnns

Datascientest. "Memoria a corto y largo plazo en LSTM." Descripción del funcionamiento de las redes LSTM para modelar secuencias temporales, aplicadas al análisis musical. https://datascientest.com/es/memoria-a-largo-plazo-a-corto-plazo-lstm

Derivative. "Getting Started With TouchDesigner." Documentación oficial para el software TouchDesigner, usado para la generación visual en tiempo real en el proyecto. https://docs.derivative.ca/

ANEXO I: ALINEACIÓN DEL PROYECTO CON LOS ODS

Este Trabajo de Fin de Grado se alinea con varios de los Objetivos de Desarrollo Sostenible (ODS) establecidos por la Agenda 2030 de las Naciones Unidas, especialmente en lo relativo a la educación, la innovación tecnológica y el acceso abierto a herramientas creativas basadas en inteligencia artificial. A través del desarrollo de un sistema modular que permite la generación automatizada de visuales reactivas en espectáculos en vivo, el proyecto contribuye de manera activa a los siguientes ODS:

ODS 4: Educación de calidad

El proyecto promueve un enfoque de aprendizaje interdisciplinar que combina conocimientos de ingeniería, inteligencia artificial, arte digital y música. Al desarrollar una solución funcional desde cero, el trabajo impulsa la formación en áreas clave para la transformación digital, como el aprendizaje automático, el procesamiento de audio, el diseño de arquitecturas modulares y la integración de entornos creativos como TouchDesigner.

Además, el carácter práctico e investigativo del proyecto favorece el aprendizaje basado en proyectos (PBL), estimula el pensamiento crítico y fomenta la exploración activa, valores fundamentales para una educación de calidad centrada en competencias digitales avanzadas.

ODS 9: Industria, innovación e infraestructura

El sistema propuesto es una contribución directa a la innovación tecnológica en el ámbito de la cultura y los espectáculos en vivo. El uso de modelos de inteligencia artificial para interpretar la música en tiempo real y generar visuales adaptativas representa una aplicación novedosa de tecnologías de deep learning fuera del entorno puramente técnico o empresarial.

Además, el proyecto apuesta por una infraestructura de software escalable y accesible, capaz de adaptarse a diferentes contextos escénicos con requerimientos mínimos. Esto favorece la adopción de nuevas herramientas creativas en el sector cultural, estimula la innovación en las industrias creativas y contribuye al desarrollo de soluciones tecnológicas más versátiles, eficientes y automatizadas.

ANEXO II

ANEXO LEGAL: Justificación del Uso de Obras Protegidas por Derechos de Autor

En el marco del desarrollo del presente Trabajo de Fin de Grado (TFG), titulado **"Implementación de la Inteligencia Artificial para la generación de visuales dinámicas en conciertos de música electrónica"**, se ha llevado a cabo el entrenamiento de modelos de inteligencia artificial con fragmentos de obras musicales que podrían estar protegidas por derechos de autor.

1. Finalidad del uso

El uso de dichos fragmentos tiene una finalidad estrictamente académica, investigadora y no comercial. Los modelos desarrollados no reproducen, distribuyen ni comunican públicamente las obras originales, sino que extraen características técnicas y patrones sonoros para realizar tareas de clasificación y análisis automatizado. En ningún caso se pretende sustituir la obra original ni vulnerar los derechos de los titulares.

2. Base legal

El uso de este tipo de materiales se encuentra amparado por el derecho a la cita y las excepciones para uso con fines docentes e investigativos**,** reconocidas tanto por la legislación española como por la normativa europea:

- **Artículo 32 del Real Decreto Legislativo 1/1996**, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, permite el uso de obras ajenas con fines de enseñanza o investigación científica, siempre que se trate de fragmentos y no se atente contra la normal explotación de la obra ni se cause un perjuicio injustificado al titular de los derechos.

- La **Directiva 2001/29/CE** de la Unión Europea establece excepciones y limitaciones para fines de investigación científica y enseñanza.

3. Medidas adoptadas

- Se ha limitado el uso a fragmentos breves de audio, necesarios y proporcionales para el entrenamiento técnico del modelo.
- No se han distribuido públicamente los archivos de audio utilizados, ni los modelos entrenados con ellos en plataformas comerciales.
- Se han respetado, en todo momento, los principios de necesidad, proporcionalidad y mínima intervención en relación con los derechos de los titulares.