UNIVERSIDAD AUTÓNOMA METROPOLITANA
UNIDAD IZTAPALAPA
DIVISIÓN DE CIENCIA BÁSICAS E INGENIERÍA
POSGRADO EN CIENCIAS
(CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN)

"GENERACIÓN DE SOLOS DE BATERÍA DE JAZZ CON HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL"

TESIS PARA OBTENER EL GRADO DE MAESTRO EN CIENCIAS
(CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN)

Presenta
ING. RODRIGO ALVARADO CASTAÑEDA
Matrícula: 2232800209

Director: DR. PEDRO LARA VELÁZQUEZ
Co-director: DRA. MARIEL ALFARO PONCE

JURADO:
Presidente: DR. ALFONSO PRIETO GUERRERO
Secretario: DR. PEDRO LARA VELÁZQUEZ
Vocal: DR. ELI GABRIEL AVIÑA BRAVO

IZTAPALAPA, CIUDAD DE MÉXICO, 3 DE OCTUBRE DE 2025

Índice general

Índice de figuras

Índice de cuadros

Resumen

En esta investigación se aborda la generación automática de solos de batería de jazz utilizando modelos Transformer, un campo poco explorado, donde tradicionalmente la batería ha ocupado un rol secundario. El estudio se centra en desarrollar y evaluar un modelo capaz de emular solos rítmicos propios del jazz. Para ello, se crearon dos bases de datos: una con 200 solos de batería en formato MIDI, representativos de distintas épocas y estilos del jazz, y otra sintética ampliada a 2000 entradas mediante modificaciones algorítmicas. Estas bases de datos, disponibles públicamente, constituyen un recurso valioso para futuras investigaciones en generación musical y análisis computacional.

El modelo implementado, basado en DistilGPT-2, fue entrenado con diferentes configuraciones de tamaño de token, 500 y 1000, y épocas de entrenamiento; 6, 9 y 12. Los resultados mostraron que la configuración con 500 tokens y 6 épocas alcanzó el mejor rendimiento, con un 50 % de eficacia en la generación de solos clasificados como auténticos por evaluadores expertos. Sin embargo, la evaluación cualitativa reveló limitaciones en aspectos como la coherencia estructural y la expresividad musical, donde los solos humanos superaron a los generados en estilo y coherencia. En contraste, los solos generados destacaron en originalidad, sugiriendo que la inteligencia artificial puede aportar ideas novedosas, aunque con menor cohesión estilística.

Las principales limitaciones identificadas incluyen la distribución desigual de los datos históricos, las restricciones del modelo para capturar matices expresivos y las limitaciones en los protocolos de evaluación. Como trabajo futuro, se propone expandir y equilibrar la base de datos, explorar arquitecturas híbridas que combinen Transformers con conocimiento musical explícito, y optimizar los métodos de evaluación integrando métricas computacionales de estilo. Los resultados de este estudio representan un avance significativo en la integración de la inteligencia artificial en la creación musical, aunque subrayan la complejidad de emular la riqueza artística y expresiva de los músicos profesionales.

Palabras clave: Jazz, batería, solos, Transformer, inteligencia artificial, generación musical, MIDI, evaluación cualitativa.

Abstract

This research addresses the automatic generation of jazz drum solos using Transformer models, a relatively unexplored field where the drum has traditionally played a secondary role. The study focuses on developing and evaluating a model capable of emulating rhythmic solos characteristic of jazz. To this end, two datasets were created: one comprising 200 drum solos in MIDI format, representative of different jazz eras and styles, and another synthetic dataset expanded to 2000 entries through algorithmic modifications. These publicly available datasets constitute a valuable resource for future research in musical generation and computational analysis.

The implemented model, based on DistilGPT-2, was trained with different configurations of token size (500 and 1000) and training epochs (6, 9, and 12). The results showed that the configuration with 500 tokens and 6 epochs achieved the best performance, with a 50% efficacy in generating solos classified as authentic by expert evaluators. However, qualitative evaluation revealed limitations in aspects such as structural coherence and musical expressiveness, where human-generated solos outperformed the AI-generated ones in style and consistency. In contrast, the generated solos stood out in originality, suggesting that artificial intelligence can contribute novel ideas, albeit with less stylistic cohesion.

The main identified limitations include the uneven distribution of historical data, the model's constraints in capturing expressive nuances, and the limitations in evaluation protocols. As future work, it is proposed to expand and balance the dataset, explore hybrid architectures combining Transformers with explicit musical knowledge, and optimize evaluation methods by integrating computational style metrics. The results of this study represent a significant advancement in the integration of artificial intelligence into musical creation, while underscoring the complexity of emulating the artistic and expressive richness of professional musicians.

Keywords: Jazz, drums, solos, Transformer, AI.

Capítulo 1

Introducción

El jazz es un estilo musical que tuvo su origen en los Estados Unidos, desarrollándose primordialmente hacia el final del siglo XIX y al principio del siglo XX, especialmente en el área del delta del río Misisipi. Nueva Orleans fue una de las ciudades clave para su evolución. Este género es conocido por su inconfundible ritmo con subdivisiones en tres notas, denominado *swing*, que resulta de la fusión de ritmos binarios y ternarios. Además, el jazz se caracteriza por sus ritmos sincopados, armonías complejas y por incluir secciones de improvisación donde los músicos crean espontáneamente en el momento, tomando como base estructuras armónicas y rítmicas predefinidas. Entre los instrumentos que más representan al jazz se encuentran la guitarra, el saxofón, el piano, el contrabajo y la batería.

El conjunto de instrumentos de percusión conocido como batería incluye una considerable variedad, tales como tarola, bombos, toms y platillos, entre otros. En sus inicios en el ámbito del jazz, la batería era percibida exclusivamente como un instrumento acompañante, encargado de mantener el ritmo y el tempo junto al contrabajo y el piano. Sin embargo, en la década de 1950, esta percepción cambió notablemente debido a la influencia del baterista y compositor Max Roach [1]. A pesar de que la batería sigue siendo el núcleo rítmico del grupo, Roach elevó su presencia en el conjunto, destacándola también como un instrumento con capacidades melódicas. Así, desarrolló secciones dedicadas a solos de batería, permitiendo una interacción similar a la de otros instrumentos. Los solos de batería en el jazz se convierten en una forma de expresión tanto rítmica como melódica, donde los músicos incorporan diversos elementos en su ejecución, tales como rudimentos, polirritmias e interacción con los diferentes instrumentos [1].

El estudio de la generación de música utilizando herramientas de inteligencia artificial (IA) no es un fenómeno reciente; de hecho, ha sido objeto de investigación durante varias décadas. Originalmente, se empleaban sistemas basados en reglas, como los autómatas y algunos algoritmos que utilizaban cadenas de Markov, junto a algoritmos evolutivos y genéticos para crear música [2]. Sin embargo, fue la introducción de las redes neuronales lo que realmente impulsó la evolución en este campo. En particular, técnicas como las Redes Neuronales Recurrentes, y específicamente las redes Long Short-Term Memory (LSTM), marcaron un punto de inflexión [2]. Estas redes tienen la capacidad distintiva a las redes neuronales tradicionales de mantener una "memoria", lo que les permite "aprender" dependencias temporales, aunque su utilidad está condicionada por la capacidad de procesamiento de los sistemas computacionales en uso [3]. Con un aumento en el poder de procesamiento y la disponibilidad de extensas bases de datos, han surgido avances significativos en el aprendizaje profundo, destacándose las redes generativas antagónicas (del inglés Generative Adversarial Networks, GANs) y los Transformers en el área de la generación de música [3]. Estas avanzadas arquitecturas pueden analizar y mejorar las conexiones entre dependencias a largo plazo y las nuevas secuencias generadas, generando resultados que presentan una musicalidad significativamente lograda y mejorada en comparación con los diferentes algoritmos de aprendizaje previamente utilizados.

La noción de "musicalidad" está estrechamente relacionada con la habilidad de un músico para transmitir sentimientos, ideas, emociones y creatividad de manera efectiva a través del medio musical. Esto incluye elementos como la expresividad, la creatividad y la sensibilidad, que son componentes esenciales de dicha capacidad. La musicalidad, intrínsecamente ligada a la experiencia humana y las diversas manifestaciones culturales, varía significativamente en su percepción. Por ejemplo, en lo que respecta al instrumento de la batería, un solo no se limita a una simple sucesión de notas, sino que representa una forma de expresión individual y personal por parte del músico. Esta característica es altamente subjetiva, dado que cada individuo la valora de manera distinta. De hecho, no hay un sistema o método para cuantificar la musicalidad en una pieza musical; esta puede encontrarse en una composición tan elaborada como una ópera, así como en algo tan aparentemente simple como un solo de batería.

En esta tesis se propone el diseño de un modelo basado en redes neuronales tipo Transformers para generar solos de batería de jazz. El estudio contempla la creación de una base de datos de solos y el desarrollo de un modelo generativo que preserve las características distintivas del género, sometiendo sus resultados a validación con músicos expertos para evaluar su musicalidad y autenticidad estilística. Si bien el proyecto no aborda explícitamente las implicaciones éticas en su diseño técnico, sí las considera como marco crítico de referencia. La literatura señala que los modelos generativos como el aquí propuesto plantean desafíos éticos sustanciales, como la atribución de autoría, el riesgo de infracción de derechos de autor por la imitación y los sesgos culturales inherentes a los datos de entrenamiento [4, 5]. Además, la evaluación con expertos contribuye a discernir la autenticidad emocional y creativa de las salidas generadas, un aspecto crucial ante la posibilidad de que la inteligencia artificial genere contenido que simplemente replique obras previas sin aportar originalidad genuina [4]. Así, aunque el enfoque del proyecto es técnico y musical, se reconoce la necesidad de desarrollar estas tecnologías de manera consciente y contextualizada en el debate ético actual.

1.1. Planteamiento del problema

La aplicación de algoritmos de IA para la generación automática de música constituye un campo de investigación ampliamente explorado en los últimos años. Se han desarrollado sistemas capaces de generar automáticamente melodías, armonías e incluso composiciones completas que emulan el estilo de grandes compositores, como Johann Sebastian Bach o Ludwig van Beethoven, donde se destacan evaluaciones por usuarios de diferentes niveles musicales, obteniendo un 50 % de rendimiento entre música generada por inteligencia artificial y música compuesta originalmente.

En el ámbito específico de la percusión, destacan los trabajos de [6] y [7], los cuales se centran en la creación automática de ritmos de batería para una diversidad de géneros musicales. Por un lado, Hutchings propuso un modelo de secuencia a secuencia, entrenado con un corpus de más de 250 partituras en compás 4/4 de estilos como pop, rock, funk y afrocubano. Su mejor técnica de muestreo logró una tasa de aceptación de los evaluadores del 39 %, aunque los resultados mostraron una variabilidad significativa dependiendo del género musical. Por otro lado, Makris et al. [7] desarrollaron un sistema de aprendizaje neuronal condicional que combina redes de memoria a corto-largo plazos y redes neuronales feed-forward. Este enfoque genera ritmos bajo restricciones métricas y armónicas, demostrando una notable capacidad para adaptarse a estilos tan diversos como disco, blues rock y metal progresivo, con resultados consistentes en la imitación estilística.

No obstante, uno de los desafíos más significativos en el desarrollo de algoritmos de IA para la improvisación de solos de batería de jazz es la ausencia de bases de datos específicas y extensivas que contengan ejemplos de solos de jazz. Mientras que hay muchas grabaciones y transcripciones disponibles, no existe una base de datos estandarizada y accesible para el

entrenamiento de modelos de IA en este dominio específico. Esta carencia dificulta la creación de modelos precisos y representativos que puedan capturar la complejidad y la creatividad inherente a los solos de batería en el jazz.

Sin embargo, la capacidad de los algoritmos de IA para replicar procesos creativos humanos, como la improvisación en el jazz, no ha sido extensivamente explorada. Este estudio tiene como objetivo diseñar y validar un algoritmo capaz de generar solos de batería de jazz, evaluando su autenticidad y calidad mediante la percepción de expertos. La investigación no solo pretende avanzar en el campo de la IA aplicada a la música, sino también aportar conocimientos sobre la naturaleza de la creatividad musical y su replicación mediante máquinas, lo que podría revolucionar el desarrollo de herramientas educativas y de entretenimiento musical.

1.2. Justificación

La presente investigación surge en respuesta a la creciente exploración de aplicar la tecnología en las artes, particularmente en la música. A lo largo de los últimos años, la IA ha revolucionado numerosos campos, incluyendo el arte y la música, donde ha mostrado un potencial significativo para la creación y la interpretación. Sin embargo, la aplicación de la IA en la generación de música jazz, especialmente en la improvisación de solos de batería, sigue siendo un tópico sin explorar. Esta investigación busca abordar hasta qué punto los algoritmos de inteligencia artificial pueden no solo replicar, sino también innovar en la creación de música jazz, un género que se caracteriza por su complejidad rítmica y su naturaleza improvisada. La relevancia de este estudio es múltiple: desde una perspectiva académica, proporcionará una contribución significativa a la literatura existente en los campos de la inteligencia artificial y la música, explorando nuevas formas de integración tecnológica en la música; a nivel práctico, los resultados tienen el potencial de influir en la industria musical, ofreciendo herramientas innovadoras para músicos y compositores, y en el ámbito educativo, donde estos algoritmos podrían servir como métodos de enseñanza y aprendizaje de la música jazz. El proyecto está motivado, además, por la necesidad de entender mejor cómo la tecnología puede capturar y expandir las características únicas del jazz, donde los solos de batería no solo requieren técnica, sino también un profundo sentido del ritmo y una capacidad para improvisar que hasta ahora se ha creído exclusivamente humana; por ello, el desafío de crear un algoritmo que pueda crear estos solos ofrece una oportunidad única para avanzar en nuestro entendimiento de la creatividad y la improvisación. La metodología propuesta combina técnicas de aprendizaje y deep learning para entrenar algoritmos capaces de generar solos de batería, siendo este enfoque adecuado debido a la capacidad de estos aprendizajes para manejar grandes cantidades de datos y aprender de ejemplos complejos, y porque permite una evaluación cualitativa de la música generada, asegurando que los solos producidos mantengan la esencia y la calidad artística del jazz. Finalmente, la investigación propuesta no solo es relevante por su potencial para aportar al desarrollo tecnológico y musical, sino también por su capacidad para inspirar futuras investigaciones en la interacción entre la inteligencia artificial y otras formas de expresión artística, teniendo a largo plazo implicaciones significativas no solo para los músicos, sino también para el desarrollo de inteligencia artificial capaz de colaborar "creativamente" con humanos en diversos campos.

1.3. Motivación

Este proyecto busca explorar la aplicación de tecnologías de la información y la música, en especial el jazz, un género definido por su improvisación y complejidad rítmica. La aplicación de la IA en este campo es nueva, ofreciendo un terreno fértil para investigar su capacidad de emular y expandir la creatividad musical.

Se busca aportar al avance técnico en IA musical y profundizar en la comprensión de la improvisación como proceso creativo. Los resultados podrían beneficiar a músicos y compositores, proporcionando nuevas herramientas para la interpretación musical, y desafiar las concepciones actuales sobre la colaboración entre humanos y máquinas en la música.

1.4. Pregunta de Investigación

¿Puede un modelo generativo de IA entrenado con solos de bateristas de jazz clásico producir patrones que repliquen características estilísticas clave, según evaluación de expertos y análisis computacional?

1.5. Hipótesis

La aplicación de algoritmos basados en aprendizaje automático y aprendizaje profundo en la generación de solos de batería de jazz puede producir resultados que son indistinguibles de las interpretaciones de bateristas expertos en jazz, tanto en términos de técnica como de expresión creativa.

1.6. Objetivos

1.6.1. Objetivo general

Implementar un algoritmo basado en redes neuronales tipo Transformer, capaz de generar solos de batería de jazz que repliquen las características estilísticas del género, evaluando su calidad mediante criterios musicales y técnicos.

1.6.2. Objetivos particulares

- Generar una base de datos de 200 solos de batería de jazz.
- Implementar un modelo de redes neuronales tipo Transformer para la generación de solos de batería.
- Generar una base de datos sintética con tamaño de 2000 solos de batería.
- Validar el modelo tipo Transformer generando un conjunto de 90 solos de batería.

Capítulo 2

Marco Teórico

2.1. Música

La música es el arte de combinar tonos de manera agradable, expresiva o inteligible, mientras que el ruido se refiere a sonidos no deseados. Sin embargo, la distinción entre ruido y música no es completamente clara, ya que ambos comparten ciertas características.

Las propiedades físicas del sonido provienen de su medio de propagación, las ondas sonoras, que se caracterizan por frecuencia, intensidad, forma de onda y duración. Además, el sonido presenta características psicológicas intrínsecamente relacionadas con sus propiedades físicas. Estas características incluyen el timbre, el volumen, el tiempo y la textura sonora [8].

2.1.1. Características físicas de la música

Frecuencia

La frecuencia es el número de ciclos que ocurren por unidad de tiempo y se determina contando los valores recurrentes de una cantidad periódica. El término "ciclos por segundo" define la frecuencia de una onda, cuya unidad de medida es el Hertz (Hz) [8].

Intensidad

La intensidad del sonido es la energía sonora transmitida por unidad de tiempo a través de una unidad de área. Su unidad de medida es el ergio por segundo por centímetro cuadrado [8].

2.1.2. Forma de onda

Una onda sonora compleja está compuesta por sobretonos. La estructura armónica o de sobretonos se describe en términos de la distribución de la intensidad y las relaciones de fase de sus componentes [8].

Duración

La duración es la longitud de tiempo que un tono persiste o termina. La unidad es el segundo o algún submúltiplo de los segundos. En notación musical, la duración es indicada por el tipo de nota que será interpretada, ver figura 2.1, [8].

2.1.3. Características psicológicas de la música

Las características psicológicas de la música se pueden clasificar en cuatro categorías principales: tonal, dinámica, temporal y cualitativa. Las características tonales incluyen aspectos

como el tono, el timbre, la melodía, la armonía y todas las formas de variación del tono. Por otro lado, las características dinámicas están determinadas principalmente por el volumen. Las características temporales abarcan elementos como el tiempo, la duración, el tempo y el ritmo, mientras que las cualitativas se relacionan con el timbre o la constitución armónica del tono [8].

Tono

El tono se define como la característica sensorial asociada a la frecuencia, que determina la posición de un sonido dentro de una escala tonal. Debido a su naturaleza subjetiva, el tono puede medirse mediante diferentes pruebas perceptivas.

El tono más bajo que se puede generar corresponde a la frecuencia más baja dentro de un espectro sonoro, mientras que el tono más alto se asocia con la frecuencia más alta de ese mismo espectro [8], ese espectro sonoro para los humanos abarca desde los 20 Hz hasta los 20 kHz.

Armonía

Se refiere al sonido de dos o más notas que se escuchan de manera simultánea. En práctica también pueden abarcar ciertos casos donde las notas se tocan de manera consecutiva, si esto evoca a un acorde que sea familiar donde el oído humano puede percibirlas como si hubieran sido tocadas al mismo tiempo.

En un sentido más técnico, la armonía se centra en el sistema desarrollado de acordes y en las reglas que rigen las relaciones entre ellos, características esenciales en la música occidental. Este sistema establece que combinaciones de acordes son aceptables o no, configurando las bases de la música occidental.

La música puede entenderse como una combinación de dos dimensiones: horizontal y vertical. Los aspectos horizontales corresponden a los elementos que se desarrollan a lo largo del tiempo, como la melodía y el ritmo. Por otro lado, el aspecto vertical se refiere a la armonía, que abarca la totalidad de los sonidos que ocurren simultáneamente en un momento dado [9].

Melodía

En la música, la melodía es el aspecto estético que surge de la sucesión de diferentes tonos en el tiempo, siguiendo un orden rítmico de tono a tono.

Una línea melódica tiene varias características importantes, entre las cuales destacan el contorno, el rango y la escala. El contorno es la forma general que describe la melodía, ya sea ascendente, descendente o con movimientos característicos.

El rango de una melodía se refiere al espacio que ocupa dentro del espectro de tonos perceptibles por el oído humano. Por otro lado, la melodía se basa en una escala, que se reconoce formalmente como un sistema de tonos a partir del cual se construyen las melodías. Las escalas pueden abstraerse enumerando los tonos utilizados en orden de altura (frecuencia). Los intervalos de una escala dentro de una melodía contribuyen significativamente a definir su carácter general [10].

Volumen

El aspecto dinámico de la música depende principalmente de la intensidad del sonido generado; a esto se le conoce como Volumen. El volumen de un sonido se refiere a la magnitud de la sensación sonora que este produce sobre un escucha. La unidad de medida del volumen es el sonio, que se define como el volumen producido por un tono puro de 1000 Hz a 40 decibeles.

Ritmo

El ritmo es la organización de las notas musicales en el tiempo. En términos generales, el ritmo (del griego rhythmos, derivado de rhein, "fluir") se define como una alternancia ordenada de elementos contrastantes [11].

En una composición musical, el ritmo constituye el patrón temporal y es el elemento de mayor importancia en la música. Puede existir sin melodía, como en los ritmos de batería, pero la melodía no puede existir sin ritmo. Aunque la melodía y la armonía son elementos que trabajan en conjunto, pueden estudiarse de manera independiente; sin embargo, ambos están intrínsecamente ligados al ritmo y no pueden separarse de él. Los elementos del ritmo son:

- **Pulso:** Es una serie de golpes uniformemente espaciados que marca el tempo y actúa como la estructura base del ritmo.[12].
- **Duración Relativa:** La duración de las notas se define a partir de una nota base con una duración de 4 pulsos, que luego se subdivide siguiendo un sistema binario, como se ilustra en la figura 2.1. En el contexto musical, el sistema binario puede entenderse como un sistema diático debido a su estructura basada en divisiones duales sucesivas.

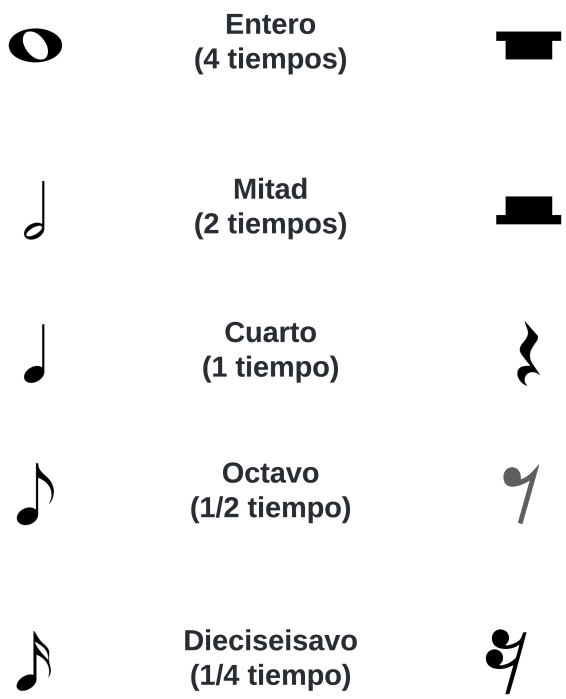


Figura 2.1. Símbolos y duración de las notas en un compás de 4/4.

- **Compás:** El compás es una unidad métrica que define la estructura rítmica de una pieza musical. Este determina cuántos tiempos tiene cada compás y cuáles se acentúan para establecer el ritmo de la composición.

  El compás se representa mediante un indicador de compás conocido coloquialmente como "quebrado musical", que se coloca al inicio de los pentagramas. El numerador del indicador señala el número de pulsos en cada compás, mientras que el denominador indica la duración de cada pulso [12]. Por ejemplo, en la figura 2.2, se muestra un indicador de compás de 4/4, eso indica que habrán 4 pulsos en cada compás y cada compás tiene una duración de una nota de cuarto, ver figura 2.1, equivalente a 1 tiempo.

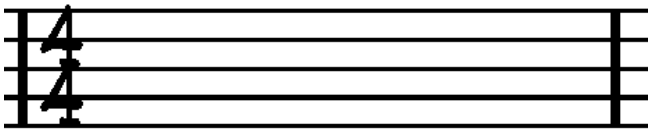


Figura 2.2. Indicador de compás de 4/4.

- **Tempo:** El tempo, también conocido como velocidad o ritmo de una pieza musical, desempeña un papel fundamental en la interpretación, actuando como el latido de la expresión musical. La palabra tempo, que significa "tiempo" en italiano, tiene su origen en el latín tempus.

  Antes del siglo XVII, la música occidental rara vez incluía indicaciones de tempo, ya que este se determinaba a partir del pulso fijo y los valores de las notas. En la era moderna, el tempo se indica principalmente de dos formas: mediante instrucciones verbales y designaciones de pulsos por minuto (del inglés beats per minute, BPM), tal como se ilustra en la tabla 2.1. Estas pueden utilizarse de forma combinada o de manera individual.

Término	**Rango Aproximado (bpm)**	**Descripción**
Largo	40–60	Muy lento.
Adagio	66–76	Lento y expresivo.
Andante	76–108	Caminar moderadamente.
Moderato	108–120	Moderado.
Allegro	120–168	Rápido y alegre.
Presto	168–200+	Muy rápido.

Cuadro 2.1. Tabla de términos de tempo musical y sus rangos en bpm.

2.2. MIDI

El protocolo de Interfaz Digital para Instrumentos Musicales (MIDI, del inglés Musical Instrument Digital Interface) es un medio de comunicación ampliamente utilizado para dispositivos musicales. Establece un conjunto de reglas muy simples para enviar diversos tipos de datos de control desde una fuente hacia un destino. MIDI admite una conexión unidireccional punto a punto, la cual puede realizarse entre equipos de hardware, software o una combinación de ambos [13].

2.2.1. Mensaje MIDI

MIDI es un lenguaje de descripción musical bajo un formato digital binario. Fue diseñado para su uso con instrumentos musicales de teclado, por lo que su estructura de mensajes está orientada a eventos de ejecución, como seleccionar una nota y luego tocarla, o ajustar parámetros típicos disponibles en teclados electrónicos [14].

Entre los mensajes MIDI más relevantes están los comandos *note-on* y *note-off*, que indican el inicio y el final de una nota, respectivamente. Cada uno de estos mensajes incluye parámetros clave como el número de nota MIDI, el valor de velocidad, el canal y una marca de tiempo. El número de nota MIDI es un entero entre 0 y 127, que codifica el tono de una nota según la escala.

La velocidad es también un entero entre 0 y 127, que controla la intensidad del sonido. En el caso de un evento *note-on*, determina el volumen, mientras que en el caso de un evento *note-off*, controla la disminución durante la fase de liberación del tono. Sin embargo, la interpretación exacta de la velocidad de la tecla depende del instrumento o sintetizador correspondiente.

Finalmente, la marca de tiempo (timestamp) es un valor entero que representa cuántos pulsos de reloj o *ticks* deben esperarse antes de que se ejecute el comando *note-on* o *note-off* respectivo [15].

La forma de comunicación del protocolo MIDI consiste en el envío de paquetes de múltiples bytes de forma serial que comienzan con un byte de estado, seguido de uno o dos bytes de dato, tal como se muestra en la figura 2.3. Los bytes son paquetes de 8 bits, donde cada bit puede tomar el valor de 0 o 1; un '1' se denomina activo y un '0', inactivo. Los bytes de estado comienzan con un '1' en su bit más significativo, por ejemplo, 1xxx xxxx, por lo que su valor decimal es 128 o mayor. Los bytes de datos comienzan con un '0', por ejemplo, 0xxx xxxx, y su valor decimal es 127 o menor. Los mensajes MIDI se dividen en dos clases principales: de Canal y de Sistema. Es de esta manera en que una partitura puede ser representada de forma digital.

Un ejemplo de la estructura de un mensaje MIDI se presenta en la Figura 2.3. El primer byte corresponde al byte de estado, identificado por tener su bit más significativo en '1'. Los bits 6 a 4 de este byte especifican el tipo de mensaje; en este caso, se trata de un mensaje de Note On, que indica la activación de una nota. Los bits 3 a 0 definen el canal MIDI por el cual se transmite el mensaje, que en este ejemplo es el canal 3.

El segundo byte es un byte de datos, reconocible por comenzar con un '0' en su bit más significativo. Este byte contiene el valor numérico de la nota musical. El valor binario '01000101' corresponde al decimal 69, que en la especificación MIDI representa la nota A4.

El tercer byte también es un byte de datos, cuyo contenido indica la velocidad de ejecución de la nota, con un valor de 100. Este parámetro controla la intensidad sonora, típicamente relacionada con la dinámica de la interpretación.

Este conjunto de tres bytes constituye un mensaje MIDI completo para la activación de una nota, transmitiendo de manera eficiente la acción musical (*"note-on"*), la altura tonal (A4) y la expresión (velocidad) a través del canal especificado.

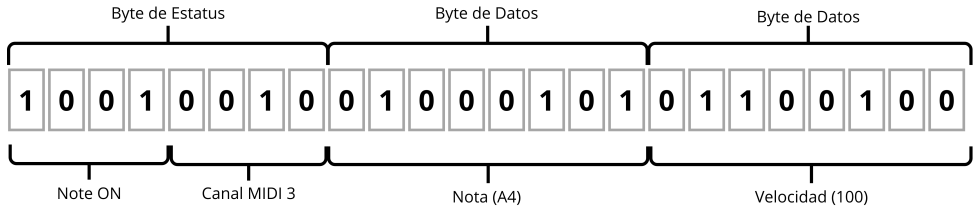


Figura 2.3. Estructura de un mensaje MIDI Note ON: byte de estatus, byte de datos (nota) y byte de datos (velocidad). Ejemplo: Canal 3, nota A4, velocidad 100.

2.2.2. MIDI como formato simbólico

El estándar original de MIDI fue ampliado para incluir la especificación del Archivo MIDI Estándar (del inglés Standard MIDI File, SMF), que define cómo almacenar datos MIDI en una computadora. Este formato permite el intercambio de datos MIDI de manera independiente al sistema operativo, sirviendo como base para la distribución eficiente de datos musicales [15].

Los archivos MIDI contienen uno o más flujos de datos con información temporal para cada evento. Este formato admite estructuras de canción, secuencia y pista, así como información de tempo y compás. Los nombres de pista y otra información descriptiva pueden almacenarse

junto con los datos MIDI. El formato permite múltiples pistas y secuencias, facilitando la transferencia de archivos entre aplicaciones que admiten múltiples pistas, garantizando así la interoperabilidad entre diferentes programas [15].

2.2.3. Evento REMI

Un evento derivado de MIDI reestructurado (del inglés REvamped MIDI-derived events, REMI) es una representación de eventos propuesta para convertir partituras MIDI en tokens discretos análogos a texto. A diferencia de las representaciones basadas en MIDI utilizadas en modelos existentes de composición musical con Transformer, REMI incorpora contexto métrico que permite modelar patrones rítmicos de manera más efectiva [16].

La representación REMI se diferencia significativamente de la representación MIDI convencional al incorporar eventos explícitos que reflejan la estructura métrica y armónica de la música. En lugar de utilizar eventos como *note-off* o *note-on*, REMI emplea eventos de duración de nota temporal que especifican directamente la longitud de cada nota, facilitando el modelado rítmico y evitando problemas de sincronización entre el inicio y el fin de las notas. Además, reemplaza los eventos *Time-Shift* que indican avances temporales relativos, por una combinación de eventos de Compás y Posición, los cuales proporcionan una cuadrícula métrica explícita que organiza la música en pulsos y compases. Esta estructura permite al modelo comprender y generar patrones rítmicos más regulares y coherentes. Adicionalmente, REMI incluye eventos de tempo para capturar cambios locales en la velocidad musical, y eventos de acorde para hacer explícita la progresión armónica, lo que brinda mayor control y expresividad en la generación musical [16].

2.3. Música Jazz

El jazz es un estilo musical que nació en Estados Unidos a finales del siglo XIX, surgido en las comunidades afroamericanas del sureste del país, con Nueva Orleans como su epicentro. Este género se caracteriza por la fusión de tradiciones musicales de África occidental, Europa y Norteamérica, integrando ritmos afroamericanos —tanto religiosos como seculares— con la instrumentación y armonías de las orquestas estadounidenses de la época, así como con influencias europeas.

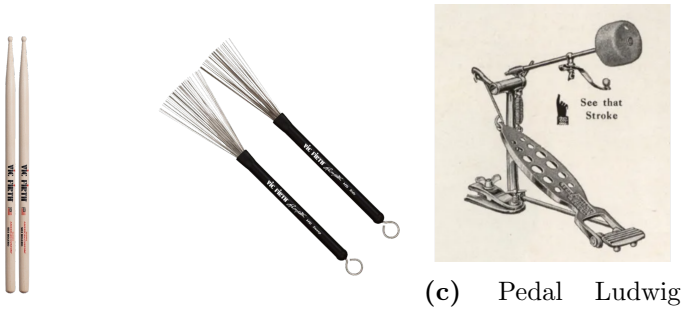
El origen del término jazz es incierto y ha sido objeto de debate. Según Walter Kingsley, podría derivar de variantes como *jas, jass, jaz, jasz o jascz*, vinculadas a la Costa de Oro africana, donde eran de uso común. Otra teoría lo relaciona con el legado francés de Nueva Orleans, sugiriendo que proviene del verbo *jaser* ("conversar animadamente"), reflejando el espíritu dinámico y colectivo de esta música.

A lo largo de su evolución, el jazz ha dado lugar a diversos subgéneros, como el ragtime, el blues, el jazz tradicional, el estilo Chicago, el swing, el boogie-woogie, el bebop, el free jazz y el jazz fusión, entre otros. A pesar de su diversidad, todos comparten elementos fundamentales: la improvisación como pilar creativo, la presencia de una sección rítmica (conformada típicamente por batería, contrabajo y piano, guitarra o banjo), melodías sincopadas, estructuras armónicas tonales con escalas de blues, y un enfoque tímbrico y técnico distintivo según cada variante [17].

2.4. La batería

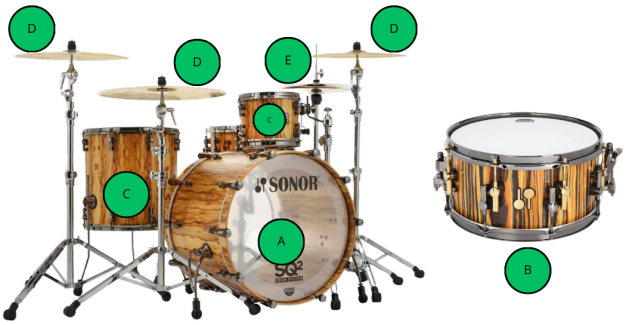
El conjunto de instrumentos conocido como batería, consiste esencialmente en un conjunto de tambores, platillos y una variedad de otros instrumentos de percusión. Estos elementos

están organizados de tal manera que un único intérprete pueda ejecutar el conjunto utilizando diferentes tipos de baquetas, como se ilustra en la figura 2.4a, así como escobillas, mostradas en la figura 2.4b, o simplemente utilizando las manos. Una configuración típica de la batería incluye componentes como el bombo, la tarola, toms, el hi-hat y platillos. No obstante, es importante señalar que los componentes específicos de la batería pueden cambiar de acuerdo a las características que presente cada género musical [18], como se representa gráficamente en la figura 2.5.



(a) Baquetas. **(b)** Escobillas. **(c)** Pedal Ludwig S.XX.

Figura 2.4. Elementos para tocar una batería.



A. Bombo B. Tarola C. Toms D. Platillos E. Hi-hats

Figura 2.5. Batería Moderna.

2.4.1. La batería en el Jazz

Los orígenes de la batería moderna se remontan a Nueva Orleans alrededor de 1900, donde surgió de la fusión entre tradiciones africanas, marchas militares y ritmos caribeños. En sus inicios, los bateristas como Baby Dodds y Zutty Singleton introdujeron los "breaks", que son breves intervenciones entre frases musicales que marcaron el nacimiento del solo de batería. Durante los años 20, el Chicago Style enfatizó los acentos en los tiempos 2 y 4, con personajes como Gene Krupa y Dave Tough refinando el uso del hi-hat y las escobillas, mientras Dee Dee Chandler sentaba las bases técnicas al crear el primer pedal de bombo artesanal en 1894.

La era dorada del jazz entre 1930-1960, consolidó a la batería como instrumento solista. Chick Webb en el Savoy Ballroom y Buddy Rich con Tommy Dorsey elevaron la técnica a nuevos niveles, siendo Rich considerado el más virtuoso de su época. El bebop de los años 40 transformó el enfoque rítmico: Kenny Clarke usó el bombo para acentos, mientras Max Roach desarrolló solos melódicos con métricas complejas utilizando compases de 5/4 o 7/4. Art Blakey incorporó polirritmos africanos en sus intervenciones con los Jazz Messengers y Elvin Jones llevó el diálogo rítmico a extremos de libertad creativa, desafiando las estructuras convencionales.

Desde los años 60, la batería rompió todas las barreras técnicas y conceptuales. Tony Williams , dominó la independencia de cuatro extremidades en sus solos, mientras el free jazz de Rashied Ali abandonó por completo las métricas tradicionales. La fusión jazz-rock de los 70, liderada por Billy Cobham y su doble bombo, incorporó setups ampliados y electrónica, sentando las bases para bateristas modernos como Vinnie Colaiuta y Dave Weckl, quienes combinaron complejidad técnica con innovación tecnológica, manteniendo sin embargo la esencia expresiva que definió al instrumento desde sus raíces en Nueva Orleans [1].

2.4.2. Notación

En el pentagrama, los instrumentos de batería tienen asignaciones específicas: cada línea y espacio designa uno diferente. Según Norman Weinberg, la tarola se ubica en el tercer espacio, el bombo en el primero y la posición de los toms depende de su número. Los hi-hats se escriben en el primer espacio bajo el pentagrama (pie) y en el espacio superior (mano). Los platillos ride y crash se colocan en la línea superior del pentagrama y en la primera línea adicional superior, respectivamente, usando cabezas de nota "X" [19].

Como se muestra en la figura 2.6, se presenta la disposición de la tarola, bombo, hihat y ride que ya se ha mencionado, y esta configuración también integra dos toms. En esta disposición, el tom 1 está asignado a la cuarta línea, mientras que el tom 2 se encuentra en la tercera línea.

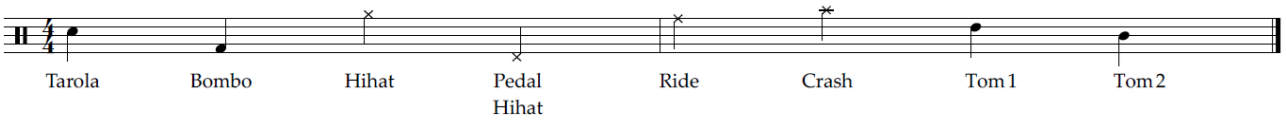


Figura 2.6. Notación estándar para batería de jazz.

Notación en el estándar MIDI

Dentro del estándar de comunicación MIDI, el canal 10 está específicamente destinado para los conjuntos de percusión, que incluyen los diversos instrumentos que conforman una batería. La tabla 2.2 presenta una lista detallada de los componentes de una batería mencionados anteriormente, junto con sus respectivos valores asignados en el estándar MIDI [14].

Nota MIDI	Instrumento
35	Bombo
38	Tarola
42	Hihat (mano)
44	Hihat (pie)
47	Tom 2
48	Tom 1
49	Crash
51	Ride

Cuadro 2.2. Asignación en el estándar MIDI de los instrumentos de una batería.

2.5. Redes Neuronales

2.5.1. Redes Neuronales Recurrentes

Las redes neuronales recurrentes (del inglés Recurrent Neural Networks,RNNs) son un tipo de redes neuronales utilizadas para el procesamiento de información secuencial [20], una de las principales características de la música. Las RNNs son una clase de modelos de aprendizaje supervisado, compuestos por neuronas artificiales con uno o más lazos de retroalimentación. El entrenamiento de una RNN de manera supervisada requiere un conjunto de datos de entrenamiento compuesto por pares de entrada-objetivo. El objetivo es minimizar la diferencia entre los pares de salida y objetivo (función de pérdida) mediante la optimización de los pesos de la red [21].

Arquitectura del modelo

Una RNN básica, figura 2.7, consta de tres capas fundamentales: entrada, oculta recurrente y salida. La capa de entrada x contiene N unidades que procesan una secuencia temporal de vectores $\{\ldots, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \ldots\}$, donde cada $\mathbf{x}_t = (x_1, x_2, \ldots, x_N)$ representa los datos en el paso t. En una configuración completamente conectada, estas unidades se vinculan a las M unidades ocultas mediante una matriz de pesos $\mathbf{W}_{IH}$.

La capa oculta mantiene un estado $\mathbf{h}_t = (h_1, h_2, \ldots, h_M)$ con conexiones recurrentes ponderadas por $\mathbf{W}_{HH}$, matriz que gobierna la transición entre estados consecutivos, $\mathbf{h}_{t-1} \rightarrow \mathbf{h}_t$. Esta dinámica define la memoria del sistema mediante:

$$\mathbf{h}_t = f_H(\mathbf{o}_t), \tag{2.1}$$

$$\mathbf{o}_t = \mathbf{W}_{IH}\mathbf{x}_t + \mathbf{W}_{HH}\mathbf{h}_{t-1} + \mathbf{b}_h, \tag{2.2}$$

donde $f_H(\cdot)$ es una función de activación que comprime el valor de $\mathbf{o}_t$ para generar el estado oculto $\mathbf{h}_t$, y $\mathbf{b}_h$ es el vector de sesgo. Las unidades ocultas están conectadas a la capa de salida mediante conexiones ponderadas $\mathbf{W}_{HO}$. La capa de salida tiene P unidades de salida $y_t = (y_1, y_2, ...y_P)$, las cuales son calculadas como:

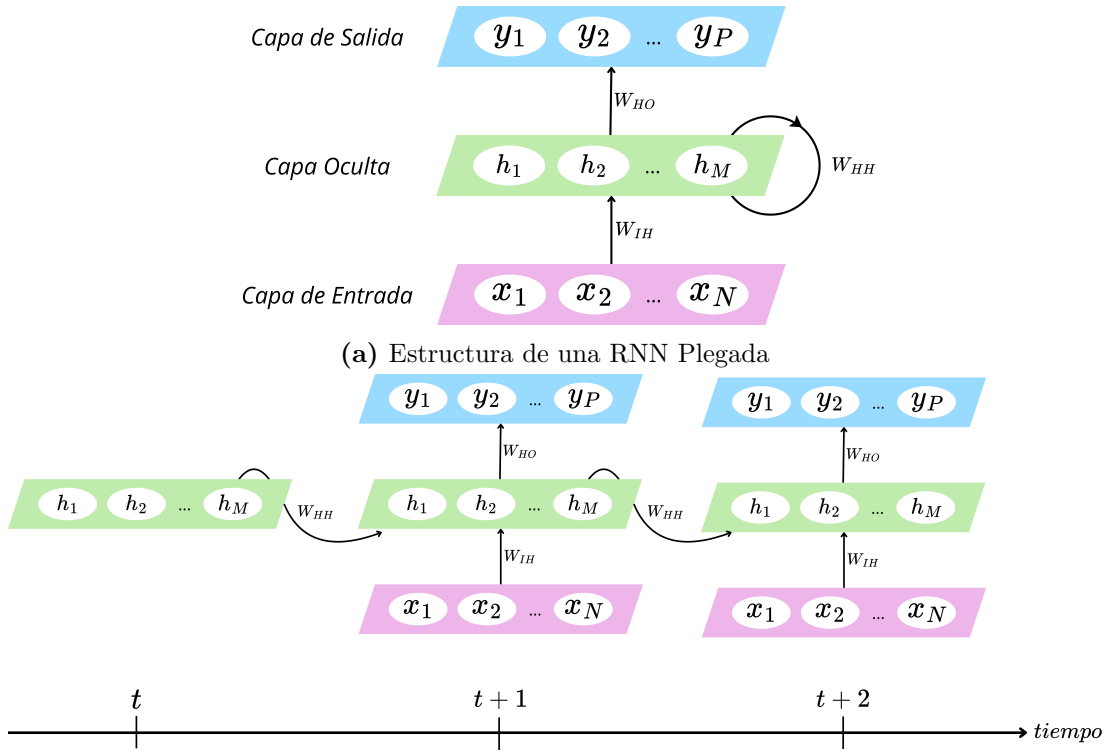
$$\mathbf{y}_t = f_O(\mathbf{W}_{HO}\mathbf{h}_t + \mathbf{b}_o), \tag{2.3}$$

siendo $f_O(\cdot)$ la función de activación de salida y $\mathbf{b}_o$ su sesgo correspondiente.Las ecuaciones 2.1 y 2.3 muestran que una RNN consiste en ciertas ecuaciones de estado no lineales, que son iterables a lo largo del tiempo. En cada paso de tiempo, los estados ocultos proporcionan una predicción en la capa de salida basada en el vector de entrada. El estado oculto de una RNN es un vector que actúa como memoria, aparte del efecto de cualquier factor externo, que resume toda la información única necesaria sobre los estados pasados de la red a lo largo de muchos pasos. [15]. [21].

Han sido diferentes aplicaciones de las RNNs en la música, se han utilizado como sistemas para clasificar las emociones generadas por escuchar música [3], también se han implementado como modelos para generar la organización de diferentes obras en listas de reproducción o en sistemas que ayuden a separar la voz de los demás elementos de la música [3], entre otras muchas aplicaciones.

Redes de Memoria a Largo-Corto Plazo

Las Redes de Memoria a Largo-Corto Plazo (del inglés Long Short-Term Memory, LSTM) constituyen un caso especial de las RNNs, diseñadas para resolver el problema del gradiente



(a) Estructura de una RNN Plegada

(b) Estructura de una RNN Desplegada a través del tiempo.

Figura 2.7. Estructura de una Red Neuronal Recurrente.

oculto en secuencias largas. Como se ilustra en la Figura 2.8, para cada instante temporal k con un vector de entrada $\mathbf{u}_k$ y N_h unidades en la capa oculta, la operación se estructura en tres puertas fundamentales. La puerta de olvido $\mathbf{f}_k \in (0,1)^{N_h}$ controla la información que se descarta de la celda de estado, calculada mediante:

$$\mathbf{f}_k = \sigma(\mathbf{W}_f \mathbf{u}_k + \mathbf{U}_f \mathbf{q}_{k-1} + \mathbf{b}_f), \tag{2.4}$$

donde σ es la función sigmoide, $\mathbf{W}_f$ y $\mathbf{U}_f$ son matrices de pesos para las conexiones de entrada y recurrentes respectivamente, $\mathbf{q}_{k-1}$ es el estado oculto anterior, y $\mathbf{b}_f$ es el vector de sesgo. Este mecanismo permite a la red adaptar dinámicamente la retención de información.

Las puertas de entrada y salida siguen un esquema similar. La puerta de entrada $\mathbf{I}_k$ determina qué nueva información se almacenará:

$$\mathbf{I}_k = \sigma\left(\mathbf{W}_I \mathbf{u}_k + \mathbf{U}_I \mathbf{q}_{k-1} + \mathbf{b}_I\right), \tag{2.5}$$

mientras que la puerta de salida $\mathbf{O}_k$ regula la información propagada al siguiente paso temporal:

$$\mathbf{O}_k = \sigma\left(\mathbf{W}_O \mathbf{u}_k + \mathbf{U}_O \mathbf{q}_{k-1} + \mathbf{b}_O\right). \tag{2.6}$$

Además, la unidad LSTM genera un candidato de actualización $\mathbf{C}_k \in (-1,1)^{N_h}$ para la celda de estado:

$$\mathbf{C}_k = \tanh\left(\mathbf{W}_C \mathbf{u}_k + \mathbf{U}_C \mathbf{q}_{k-1} + \mathbf{b}_C\right). \tag{2.7}$$

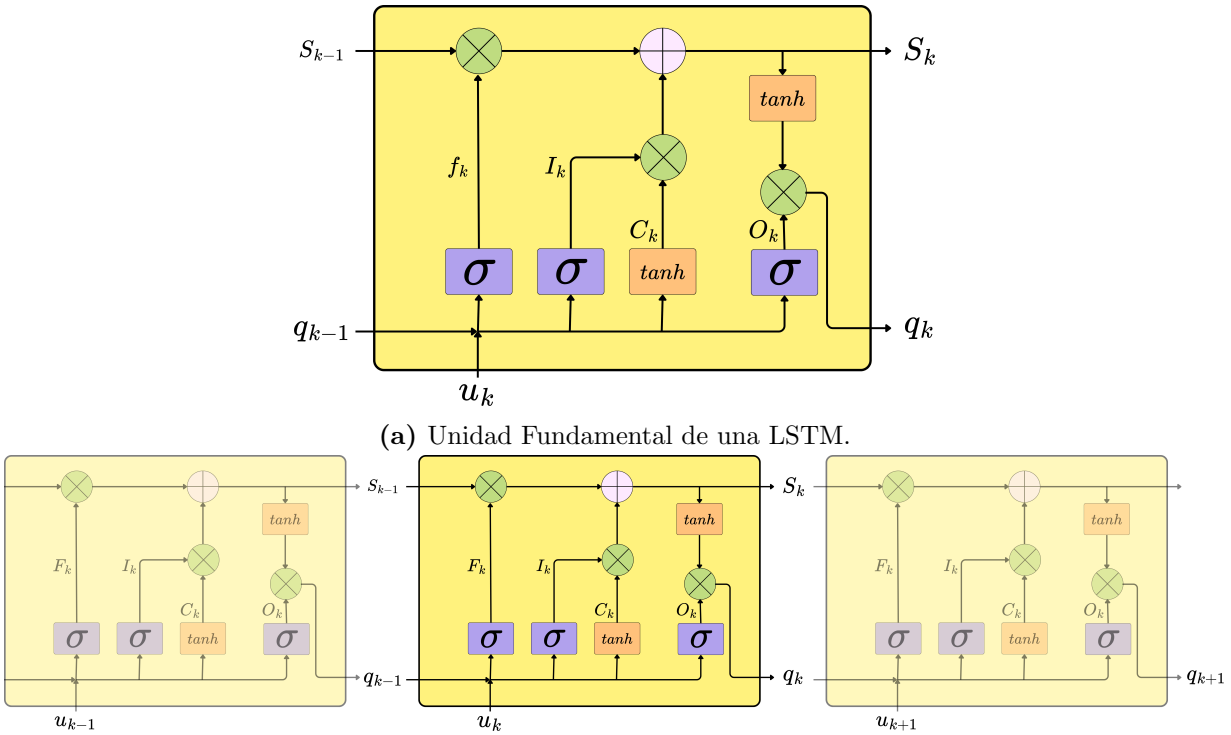
La celda de estado $\mathbf{S}_k$ se actualiza combinando estas operaciones mediante:

$$\mathbf{S}_k = \mathbf{f}_k \circ \mathbf{S}_{k-1} + \mathbf{I}_k \circ \mathbf{C}_k, \tag{2.8}$$

donde $\circ$ denota el producto de Hadamard. Finalmente, el estado oculto $\mathbf{q}_k$ se calcula como:

$$\mathbf{q}_k = \mathbf{O}_k \circ \tanh(\mathbf{S}_k), \tag{2.9}$$

completando así el flujo de información a través de la unidad LSTM. Esta arquitectura permite manejar eficientemente dependencias temporales tanto a corto como a largo plazo.



(a) Unidad Fundamental de una LSTM.

(b) Transición de Estados en una Red LSTM.

Figura 2.8. Arquitectura detallada de una celda LSTM, mostrando sus compuertas y flujo de información.

2.5.2. Transformers

Los Transformers constituyen una arquitectura de redes neuronales diseñada principalmente para la tarea de transducción de secuencias, evitando el uso de recurrencias y convoluciones. Este modelo se apoya exclusivamente en el **mecanismo de atención**, permitiendo así una mayor paralelización durante el entrenamiento y un menor costo computacional en contraste con las redes neuronales recurrentes y las redes convolucionales [22].

El concepto de atención se ha introducido como un complemento en modelos recurrentes para permitir el modelado de relaciones entre elementos sin importar su distancia. El Transformer elimina la dependencia de recurrencias, sustituyéndolas por la autoatención [22].

2.5.3. Arquitectura

La mayoría de los avanzados modelos neuronales para transducción de secuencias siguen un diseño de codificador-decodificador. Bajo este esquema, el codificador convierte la secuencia de entrada conformada por representaciones simbólicas $x = (x_1, ..., x_n)$ en una secuencia continua representada como $z = (z_1, ..., z_n)$. Posteriormente, el decodificador toma z para originar una secuencia de salida $(y_1, ..., y_m)$ al generar símbolos de manera secuencial y uno a la vez. El modelo opera de modo autoregresivo en cada uno de estos pasos [22].

El Transformer sigue esta arquitectura general, pero emplea capas de autoatención apiladas y capas completamente conectadas punto a punto, tanto en el codificador como en el decodificador, tal como se muestra en la figura 2.9.

Codificador

El codificador consta de una secuencia de $N = 6$ capas, todas ellas con la misma estructura. Cada una de estas capas incluye dos subcapas distintas:

1. Un sistema de autoatención con múltiples cabezas.
2. Una red neuronal feed-forward, totalmente conectada y aplicada de manera independiente a cada posición.

Se incorporan conexiones residuales alrededor de cada subcapa, seguidas de un proceso de normalización por capas. Así, la salida de cada subcapa se obtiene como: $LayerNorm(x + Sublayer(x))$, donde $Sublayer(x)$ es la función específica implementada por dicha subcapa [22].

Decodificador

El decodificador está estructurado con una pila de $N = 6$ capas iguales. Aparte de las dos subcapas integradas en cada capa del codificador, el decodificador añade una tercera subcapa que aplica atención multi-cabeza sobre la salida del codificador. Similar al codificador, cada subcapa está rodeada por conexiones residuales, seguidas de normalización por capas. Además, la subcapa de autoatención del decodificador se ajusta para impedir que las posiciones atiendan a aquellas posteriores. Esta técnica de enmascaramiento, junto con un desplazamiento de una posición en los embeddings de salida, asegura que las previsiones para la posición i solamente dependan de las salidas conocidas en posiciones previas a i [22].

Atención

Una función de atención puede ser definida como un proceso que efectúa la correspondencia entre una consulta y un conjunto conformado por pares clave/valor, resultando en una salida, con todos los componentes funcionando como vectores. Esta salida se obtiene mediante la suma ponderada de los valores. Los pesos correspondientes a cada uno de estos valores son determinados a través de una función de compatibilidad que evalúa la relación entre la consulta dada y su clave asociada. Hay dos clases de funciones de atención: la función **"Atención mediante producto punto escalado"** y la función **"Mecanismo de atención múltiple"**.

Atención mediante producto punto escalado

La atención mediante producto punto escalado, figura 2.10a, utiliza claves y consultas con dimensiones de d_k y valores de dimensión d_v. Esto se logra calculando el producto punto entre las consultas y los valores, normalizando dividiendo por d_k y aplicando posteriormente una función $softmax$ para asignar los pesos adecuados a los valores. En la práctica, se calcula la función de atención sobre un conjunto de consultas de manera simultánea, agrupadas en una matriz Q. Las claves y los valores también se agrupan en matrices K y V, respectivamente. El cálculo de la matriz de salidas se realiza mediante la siguiente operación:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.10}$$

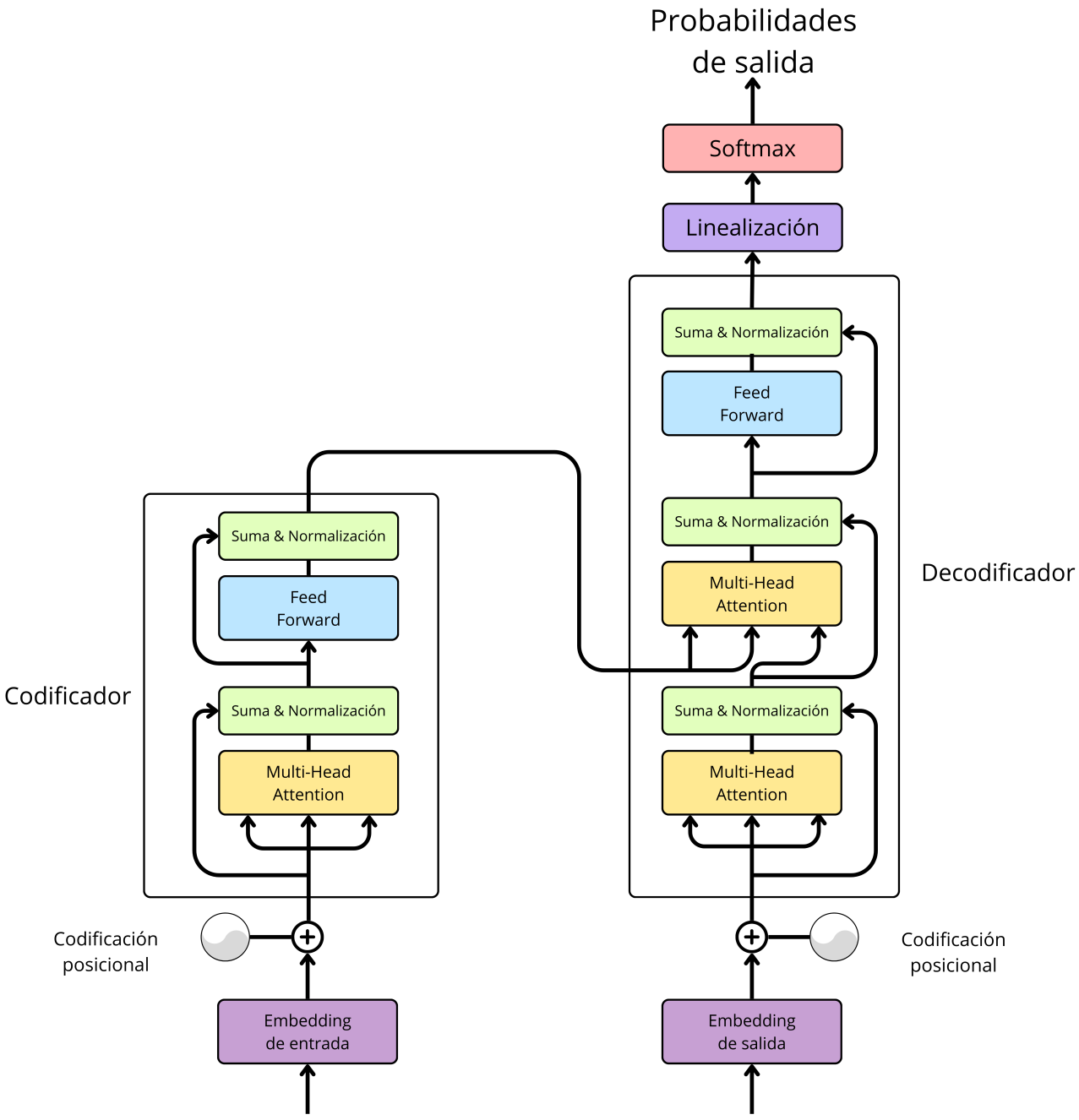


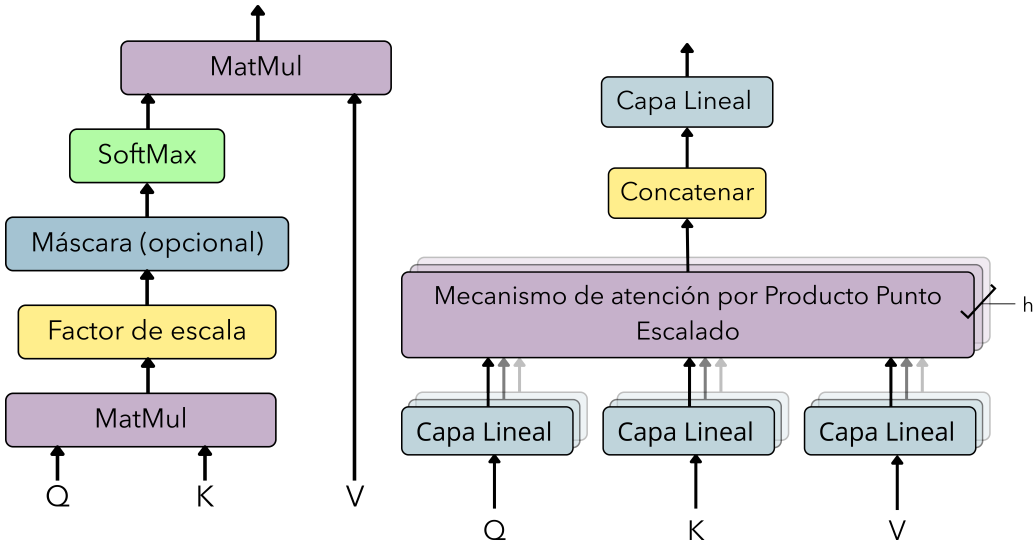
Figura 2.9. Arquitectura de un Transformer.

Mecanismo de atención múltiple

En lugar de usar una sola función de atención que emplea consultas, claves y valores con la dimensión d_{model}, el mecanismo de atención múltiple lleva a cabo proyecciones lineales h veces de las consultas, claves y valores, cada vez utilizando diferentes proyecciones lineales que son aprendidas para las dimensiones d_k, d_k y d_v, respectivamente, figura 2.10b. Posteriormente, en estas consultas, claves y valores proyectados, aplicamos la función de atención de manera simultánea, lo que genera valores de salida con una dimensión de d_v.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(head_1, head_2, \ldots, head_h)W^O \\ \text{donde } head_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.11}$$

Dónde las proyecciones son parámetros matriciales $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ y $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.



(a) Mecanismo de Atención mediante Producto Punto Escalado.

(b) Mecanismo de Atención Múltiple.

Figura 2.10. (a) Esquema del mecanismo de atención mediante producto punto escalado, mostrando las operaciones de multiplicación de matrices , normalización , máscara y factor de escala. (b) Representación del mecanismo de atención múltiple, donde Q, K y V denotan consultas, claves y valores respectivamente.

Redes Neuronales Feed-Forward por Posición

Aparte de las subcapas de atención, cada nivel tanto del codificador como del decodificador incorpora una red neuronal feed-forward completamente conectada que se aplica de manera independiente e idéntica a cada posición. Esta red neuronal consta de dos transformaciones lineales, separadas por una función de activación ReLU. La operación utilizada en esta subcapa, esta definida por la siguiente ecuación:

$$FFN(x) = \text{máx}(0, xW_1 + b_1)W_2 + b_2 \tag{2.12}$$

Donde:

- x: Tensor de entrada.
- W_1: Matriz de pesos de la primera capa lineal, con dimensiones.

- b_1: Vector de sesgo (bias).
- máx: Función de activación ReLU, que aplica un umbral a los valores. negativos, haciéndolos cero, y preserva los positivos.
- W_2: Matriz de pesos de la segunda capa lineal.
- b_2: Vector de sesgo final de dimensión.

2.5.4. Modelo de lenguaje de gran tamaño

Los modelos de lenguaje (LMs) se fundamentan en el cálculo de distribuciones probabilísticas de secuencias de símbolos discretos como palabras o tokens, con el objetivo de modelar texto para predecir coherentemente el siguiente elemento en un contexto dado. Este enfoque se basa en la factorización de la probabilidad conjunta $p(x)$ de una secuencia $(s_1, s_2, ..., s_n)$ en un producto de probabilidades condicionales, expresado formalmente como:

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, s_2, ..., s_{n-1}) \tag{2.13}$$

Lo que permite tanto el muestreo eficiente como la estimación de distribuciones completas o condicionales. La flexibilidad de este marco teórico ha permitido su evolución hacia arquitecturas más sofisticadas, destacando los modelos basados en transformers que, mediante mecanismos de autoatención, logran capturar dependencias complejas en secuencias extensas, mejorando significativamente la capacidad predictiva y la coherencia en la generación de texto [23].

GPT2

OpenAI desarrolló el modelo de lenguaje GPT-2, que está basado en la arquitectura de transformers. Este modelo cuenta con 1.5 mil millones de parámetros y fue entrenado con el conjunto de datos WebText. Dicho conjunto de datos estaba constituido por 45 millones de páginas web que fueron seleccionadas manualmente, con la finalidad de evitar textos de baja calidad y duplicados [23].

El modelo adopta un enfoque de aprendizaje cero-shot, lo que implica realizar tareas sin entrenamiento específico o ajuste fino para cada actividad. GPT-2 tiene la capacidad de ejecutar funciones tales como traducción, resumen, comprensión de lectura y responder preguntas, utilizando únicamente instrucciones en lenguaje natural como entrada. Además, GPT-2 utiliza mecanismos de autoatención para captar dependencias de largo alcance en el texto [23].

El modelo aún tiene ciertas limitaciones, ya que no aprovecha al máximo el conjunto de datos de entrenamiento. Esto sugiere la posibilidad de mejorar su desempeño mediante la implementación de modelos más grandes o el uso de bases de datos más completas y detalladas. Además, su rendimiento en tareas como la generación de resúmenes y la traducción aún es básico y queda por detrás de los métodos supervisados [23].

DistilGPT2

DistilGPT2 constituye una implementación derivada del modelo GPT-2 original mediante técnicas de destilación del conocimiento. El proceso de compresión reduce la escala del modelo a 82 millones de parámetros, conservando 12 capas de transformadores con una dimensión de embedding de 768 unidades y 12 cabezales de atención. Esta configuración permite el procesamiento eficiente de secuencias de hasta 1024 tokens mediante mecanismos de autoatención causal, preservando la capacidad generativa de lenguaje a pesar de la reducción paramétrica.

La eficiencia computacional se logra mediante la optimización de las operaciones de atención y la reducción dimensional en las capas feed-forward. El modelo emplea tokenización por Byte Pair Encoding con un vocabulario de 50257 entradas, manteniendo la compatibilidad arquitectónica con modelos de la serie GPT. Evaluaciones técnicas reportan un rendimiento del 97 % respecto al modelo original en tareas de modelado de lenguaje, con un throughput de inferencia de 21.3 muestras/segundo en hardware GPU estándar [24].

2.6. Generación automática de música

En las últimas décadas, la generación automática de música con IA ha experimentado una notable evolución. Ha pasado de utilizar métodos basados en reglas preestablecidas a adoptar modelos de aprendizaje profundo que analizan patrones en los datos musicales. Este progreso se debe a la habilidad de la IA para interactuar con varios niveles de abstracción en la música, abarcando desde la estructura básica como pueden ser las notas y los ritmos, hasta nociones más complejas como la armonía, la estructura general y la expresividad emocional [2].

2.6.1. Enfoques No Adaptativos: Reglas y Sistemas Generativos

Los sistemas iniciales de composición algorítmica basados en reglas significaron el comienzo de la generación de música asistida por computadora. Un ejemplo de este enfoque son los autómatas celulares (AC), que son capaces de generar conductas complejas a partir de reglas simples de interacción local. En estos sistemas, una cuadrícula de células se dispone de manera que cada célula modifica su estado basándose en el estado de las células vecinas, siguiendo un conjunto de reglas predefinidas. Al aplicarse al ámbito musical, esta estructura permite transformar los estados de las células en parámetros musicales tales como el tono, la duración o la intensidad, dando lugar a patrones musicales emergentes y complejos [2].

El uso de ACs en música ha resultado ser especialmente exitoso en la producción de texturas rítmicas y melódicas. Un ejemplo destacado es la obra “Horos” de Iannis Xenakis, que demuestra como estos sistemas pueden ir más allá de su base matemática para generar estructuras musicales de carácter orgánico. En esta obra, Xenakis tradujo la evolución de los estados de los CA en parámetros musicales, creando una composición donde la complejidad del sistema se refleja en desarrollos armónicos y rítmicos notablemente sorprendentes [2].

Sistemas-L, por otro lado, proporcionan un enfoque alternativo que se fundamenta en gramáticas generativas. Estos sistemas funcionan aplicando reiteradamente reglas de reescritura sobre una secuencia inicial de símbolos, de tal forma que cada símbolo puede denotar tanto eventos musicales precisos como operaciones de transformación. La fortaleza de los sistemas-L se fundamenta principalmente en su capacidad para crear estructuras análogas y jerárquicas, características que se reflejan naturalmente en muchas formas musicales convencionales [2].

Estos sistemas poseen importantes limitaciones en cuanto a su flexibilidad creativa. Su naturaleza determinista significa que todas las posibles variaciones musicales deben estar incluidas en las reglas iniciales, limitando su capacidad para ajustarse a contextos compositivos no previstos. La calidad de la música generada depende además de la habilidad del programador para prever y codificar la relación entre las estructuras matemáticas y sus equivalentes musicales. Esta dependencia del conocimiento experto humano ha impulsado la aparición de enfoques alternativos que sean capaces de aprender directamente de corpora musicales existentes, lo que representa una transición hacia sistemas más autónomos y adaptables.

2.6.2. Métodos de aprendizaje adaptativo para generación musical

Los sistemas para la generación automática de música han progresado significativamente, desarrollándose en modelos que aprenden directamente de ejemplos musicales. Los modelos probabilísticos, tales como las cadenas de Markov, fueron pioneros en mostrar resultados alentadores. Estos sistemas procesan extensas bases de datos musicales a fin de identificar patrones recurrentes, incluyendo progresiones de acordes y las interrelaciones entre melodía y armonía. Un caso destacado es el software MySong, que tiene la capacidad de crear acompañamientos armónicos para melodías vocales en una variedad de estilos [2].

A pesar de esto, dichos métodos encuentran obstáculos al tratar de entender la estructura completa de una composición musical, contemplando tanto las repeticiones como las variaciones temáticas. Esto impulsó la utilización de redes neuronales, especialmente las LSTM, las cuales tienen la habilidad de mantener patrones a través de intervalos de tiempo prolongados. Gracias a estas redes neuronales, se han producido sistemas como DeepBach, que es capaz de generar música coral en un estilo barroco con alta coherencia [2].

Los algoritmos de inteligencia artificial conocidos como redes neuronales tipo transformers se han establecido como herramientas esenciales para la creación musical. Gracias a su habilidad para analizar y prever secuencias complejas, pueden componer melodías, armonías y ritmos de manera independiente, superando la eficiencia de las redes LSTM. A diferencia de los métodos convencionales, los transformers emplean el mecanismo de autoatención para captar las relaciones y conexiones entre notas y patrones musicales a lo largo del tiempo. Esta característica los hace especialmente valiosos para generar música estructurada de manera coherente [25].

Algunos modelos, como Music Transformer y Pop Music Transformer, han demostrado que pueden generar piezas musicales largas y variadas, incluso superando a otras técnicas en evaluaciones realizadas por oyentes humanos. Otros, como Jukebox, van más allá al producir no solo instrumentales, sino también voces, adaptándose a diferentes estilos musicales. Además, permiten cierto control, como ajustar el género o la emoción de la música generada, lo que los hace útiles para compositores que buscan inspiración o herramientas creativas [25].

No obstante, el elevado uso de recursos computacionales y la demanda de una mayor diversidad de datos de entrenamiento son necesarios para incluir géneros que están menos representados, como el jazz. Sin embargo, los transformers marcan un progreso notable en la generación musical asistida por IA, lo que genera nuevas oportunidades para los artistas, los educadores y también para toda la industria musical.

Capítulo 3

Estado del Arte

La generación automática de música mediante inteligencia artificial ha experimentado avances significativos en las últimas décadas, gracias al desarrollo de modelos cada vez más sofisticados y a la disponibilidad de bases de datos musicales extensas. Este capítulo presenta una revisión sistemática del estado del arte en este campo, centrándose en los modelos y técnicas más relevantes aplicadas a la generación de música, con especial énfasis en el jazz y los solos de batería.

Esta revisión, tabla 3.1, cubre el uso de modelos basados en RNN, LSTM y Transformers en la generación de música. Se estudian las características de cada técnica empleada, junto con las estrategias aplicadas para tratar los datos musicales y para representar información simbólica. Además, se examinan los criterios utilizados para evaluar la calidad de las composiciones producidas.

Artículo	**Modelo**	**Base de datos**	**Procesamiento**	**Objetivo**	**Características**	**Validación**
Deep Learning for Expressive Music Generation, 2020 [26]	LSTM	36 archivos MIDI de Bach y grabaciones en vivo	Extracción de datos MIDI con Music21	Generar música expresiva con variaciones en velocidad y duración de notas	Tres redes LSTM independientes para notas, velocidades y duraciones	Pruebas con usuarios comparando composiciones humanas y artificiales
Automatic Composition System Based on Transformer XL, 2024 [27]	Transformer XL mejorado	Más de 700 archivos MIDI de piano popular	Conversión de MIDI a eventos usando REMI	Generar música coherente utilizando información bidireccional	Mecanismo de Mask mejorado para atención bidireccional	Indicadores objetivos, tarea de continuación musical, clasificación de piano rolls, pruebas con usuarios
Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation, 2023 [28]	Music Transformer (reimplementado)	Classical string quartets (CSQ) y classical piano improvisations (CPI)	Datos MIDI sin timing expresivo (CSQ) y con timing expresivo (CPI)	Comparar métodos de aprendizaje profundo y no profundos para generación musical	Evaluación en 6 dimensiones musicales mediante estudio de escucha	Estudio de escucha con 50 participantes y pruebas de hipótesis bayesianas no paramétricas

Automated Creation of Music using Deep Learning, 2024 [29]	LSTM	Notación ABC, esta notación representa las notas musicales con letras de la A a la G y símbolos para alteraciones, duración y clave, siendo eficaz para procesar música con modelos de deep learning [29], y MIDI	Notación ABC, convertida a MIDI y luego a MP3	Generar música automáticamente usando aprendizaje profundo	Enfoque en generación de melodías y armonías	Evaluación subjetiva de la calidad musical y consistencia de las salidas
Generating Music with Emotion Using Transformer, 2021 [30]	Transformer-XL con tokens de emoción	PEMIDI (210 archivos MIDI etiquetados manualmente)	Tokenización REMI con tokens de emoción (Emotion-Start/End)	Generar música con emociones controlables (positiva, neutral, negativa)	Integración de emociones mediante tokens, estructura musical limitada	Encuesta MOS (Mean Opinion Score) a 30 personas
Application and Research of Music Generation System Based on CVAE and Transformer-XL in Video Background Music, 2025 [31]	Transformer-XL integrado con CVAE (IUC-VAE)	Datos de música y videos	Extracción de características de movimiento (flujo óptico) y normalización	Generar música de fondo sincronizada con el ritmo y emoción de videos	Control preciso del ritmo y emoción, integración de atención mejorada	Métricas objetivas (similitud rítmica, densidad de acordes) y análisis tonal
CPS: Full-Song and Style-Conditioned Music Generation with Linear Transformer, 2022 [32]	Linear Transformer con tokens de estilo	ADL Piano MIDI (11,086 canciones de piano)	Representación CPS (Compound Word + Style) con metadatos de género	Generar canciones completas con estilos específicos (pop, clásico, jazz, rock)	Distinción entre elementos de bajo y alto nivel, estrategias adaptativas	Métricas objetivas (rango de tono, clases de tono únicas) y clasificación de género

Learning Adversarial Transformer for Symbolic Music Generation, 2023 [33]	Adversarial Transformer	907 archivos MIDI de guitarra clásica (725 entrenamiento, 90 validación, 92 prueba)	Codificación MIDI a secuencias de eventos (programa, nota, velocidad, tiempo). Aumento de datos con cambios de tono y tiempo.	Generar música simbólica larga con alta musicalidad.	Combina GAN con Transformer. Recompensa por paso (REGS) basada en armonía local y global.	Métricas objetivas (repeticiones de notas, duración) y evaluación humana (6 dimensiones musicales).
Generating Music With Emotions , 2023 [34]	Encoder-Decoder basado en GRU / Transformer	11,528 archivos MIDI con letras (LMD-full y reddit). 170k segmentos etiquetados automáticamente.)	Representación música como pares sílaba-nota. Etiquetado automático de emociones usando BERT fine-tuned.	Generar letra y melodía condicionadas a emociones (positivas/negativas).	Clasificador de emociones y búsqueda por haz emocional (EBS). Skip-gram para embeddings.	Evaluación objetiva (distribuciones de notas) y subjetiva (20 participantes, calidad y emoción percibida).
Theme Transformer: Symbolic Music Generation With Theme-Conditioned Transformer, 2023 [35]	Theme Transformer (encoder-decoder con atención paralela y codificación posicional alineada a temas)	POP909 (909 canciones de pop mandarín)	Segmentación en fragmentos de 2 compases, cuantización de notas, filtrado por firma de tiempo 4/4	Generar letra y melodía condicionadas a emociones (positivas/negativas).	Uso de aprendizaje contrastivo para extraer temas, módulo de atención paralela con compuertas XOR.	Evaluación objetiva (métricos de coherencia y temas) y subjetiva (test de escucha con usuarios).
Acapella-based music generation with sequential models utilizing discrete cosine transform, 2024 [36]	Autoencoder + GRU	250 voces humanas y 250 sonidos de guitarra (extraídos de YouTube).	Normalización de datos, segmentación en fragmentos cortos o transformada DCT (Discrete Cosine Transform)	Generar instrumentos musicales a partir de voces humanas (a capela).	Tres esquemas de entrada: corto, peinado y DCT; uso de autoencoders para compresión.	Evaluación de pérdida (MAE) y prueba auditiva para calidad y armonía
An automatic music generation method based on RSCLN Transformer network, 2024[37]	Hyperbolic Music Transformer	POP909	Representación REMI	Generar música estructurada con jerarquía	Atención hiperbólica, modelado en espacio hiperbólico, optimización con RSGD	Evaluación objetiva: Tasa de Información (IR) y MGEval (métricas de similitud). Evaluación subjetiva: Encuestas a participantes.

Hyperbolic Music Transformer for Structured Music Generation, 2023 [38]	RSCLN Transformer-XL	775 archivos MIDI de piano	Conversión a eventos REMI, aumento de datos	Generar música con estructura similar a la original, evitando desaparición de gradientes	Uso de REMI para representación rítmica, multi-head attention, normalización recursiva	Evaluación objetiva (entropía de tono, similitud rítmica) y subjetiva (test de escucha).

Cuadro 3.1. Estado del Arte.

Tras el análisis de los trabajos presentados en el estado del arte, se puede concluir que si bien existe una diversidad de enfoques y arquitecturas, desde modelos basados en LSTM y unidad recurrente cerrada (del inglés Gated Recurrent Unit,GRU) hasta propuestas más recientes como los Transformers, estos últimos han surgido como la tecnología predominante y más prometedora en los últimos años. Los Transformers, especialmente en variantes como Transformer XL o con mecanismos de atención mejorados, demuestran una capacidad superior para capturar dependencias a largo plazo y generar composiciones musicales coherentes y estructuralmente ricas. Sin embargo, la comparación directa entre diferentes técnicas resulta compleja debido a la subjetividad de la evaluación musical, la variedad de bases de datos y las representaciones simbólicas utilizadas. Las métricas de evaluación van desde el uso de medidas objetivas, como la similitud rítmica o entropía tonal, hasta subjetivas, como son las evaluaciones por usuarios, lo que dificulta establecer una superioridad absoluta de un modelo sobre otro.

En el ámbito específico de la generación automática de solos de jazz para batería, este proyecto introduce un enfoque innovador al tratar el instrumento no solo como un elemento rítmico sino también melódico, una perspectiva fundamental dentro de la tradición jazzística. Mientras la mayoría de los trabajos existentes se concentran en patrones de acompañamiento, grooves o estructuras rítmicas predecibles, esta investigación explora la capacidad de la batería para asumir un rol protagónico en la improvisación. Este enfoque amplía las posibilidades de la inteligencia artificial en la composición musical al abordar un vacío evidente en la literatura especializada.

Capítulo 4

Metodología

En este capítulo se explica el diseño experimental y los métodos utilizados para desarrollar un modelo basado en redes neuronales de tipo Transformer que tiene la capacidad de generar solos de batería en el estilo jazz. La metodología se organiza en cuatro etapas fundamentales para su implementación: en primer lugar, se lleva a cabo la creación y extensión de una base de datos compuesta por solos de batería; en segundo lugar, se procede a la implementación y posterior entrenamiento del modelo propuesto; en tercer lugar, se inicia la generación de secuencias musicales utilizando el modelo; y finalmente, se realiza una validación cualitativa donde expertos evalúan el resultado.

Se detallan los recursos computacionales utilizados, como Google Colab Pro con GPU NVIDIA Tesla T4, y las herramientas de software, incluyendo librerías especializadas en procesamiento de datos MIDI y modelos de lenguaje. Asimismo, se explican los criterios para la selección y síntesis de los datos, el preprocesamiento de archivos musicales, y la configuración del modelo DistilGPT2 para adaptarlo a la generación de patrones rítmicos y melódicos propios del jazz.

Finalmente, se presenta el protocolo de evaluación, que incluye la participación de bateristas expertos para valorar la autenticidad, coherencia y viabilidad técnica de los solos generados. Este enfoque integral busca no solo demostrar la eficacia del modelo propuesto, sino también establecer un marco reproducible para futuras investigaciones en generación musical asistida por inteligencia artificial.

4.1. Diseño experimental

Para la ejecución del diseño experimental se empleó la plataforma de computación en la nube Google Colab Pro. Esta herramienta proporcionó acceso a una unidad de procesamiento gráfico (GPU) NVIDIA Tesla T4 con 16 GB de memoria RAM y 2560 núcleos CUDA. El procesamiento central fue gestionado por un CPU Intel Xeon de doble núcleo con una frecuencia de 2.20 GHz, complementado con 25 GB de memoria RAM, recursos sujetos a la disponibilidad dinámica de la plataforma. El almacenamiento masivo se implementó mediante Google Drive con una capacidad de 2 TB.

La selección de este entorno se sustentó en un análisis comparativo de rendimiento frente a una configuración local compuesta por 16 GB de RAM, un procesador AMD Ryzen 5 a 4.2 GHz y una GPU NVIDIA RTX 3060. Las ventajas técnicas de la plataforma en la nube incluyeron: una reducción significativa en los tiempos de ejecución para el entrenamiento de modelos, escalabilidad optimizada para el manejo de grandes volúmenes de datos, y la eliminación de conflictos de compatibilidad hardware-software. Cabe destacar que la experiencia previa en el uso de la plataforma permitió una implementación eficiente sin requerir tiempos adicionales de adaptación.

Entre las limitaciones identificadas se registró la interrupción automática de sesiones tras períodos de inactividad y la dependencia de la asignación dinámica de recursos. No obstante, el balance costo-beneficio confirmó la superioridad técnica de la plataforma para los requerimientos específicos de procesamiento de datos y entrenamiento de modelos en este estudio.

El desarrollo experimental fue realizado utilizando Python, comenzando con la versión 3.8, debido a su robusta versatilidad y amplia aceptación en el ámbito de la inteligencia artificial. Este lenguaje proporciona un ecosistema abundante en librerías especializadas. Python permitió la fácil integración de librerías como **pretty_MIDI** [39], **MIDItok** [40], así como las librerías presentes en la plataforma **Hugging Face**, específicamente **Transformers** [41] y **datasets** [42].

La librería **Transformers** incluye implementaciones de vanguardia de modelos de lenguaje, como GPT-2 y BERT, entre otros, y proporciona herramientas para entrenamiento, ajuste fino e inferencia. Por otro lado, la librería **datasets** se especializa en el manejo eficiente de la carga, procesamiento y gestión de conjuntos de datos, estando especialmente optimizada para aplicaciones de procesamiento del lenguaje natural.

4.1.1. Generación de la base de datos de 200 solos de batería de jazz

Para la creación de esta base de datos, se tomó en cuenta una selección de solos icónicos y emblemáticos dentro del género del jazz, abarcando un periodo que va desde 1938 hasta 2018. Esta selección se basó en diversos subgéneros de jazz que derivan del **swing**, tales como el hardbop, el bebop, el jazz tradicional, entre otros. La base de datos recopila solos de destacados músicos, incluyendo a Max Roach, Charlie Parker, Elvin Jones, Philly Joe Jones, Joe Morello, Gene Krupa, Buddy Rich, entre otros reconocidos exponentes del género.

Los solos de batería de jazz fueron transcritos usando el programa de notación musical Sibelius. En ese mismo software, se generaron y exportaron los archivos MIDI que forman la base de datos, accesible en `https://github.com/rodrums21/DrumTransformer-`.

4.1.2. Generación de una base de datos sintética con tamaño de 2000 solos de batería

Para complementar la base de datos original de 200 solos y garantizar una mayor diversidad en los patrones melódicos y tímbricos, se desarrolló una metodología de aumento de datos que permitió expandir el conjunto a 2000 muestras. Este proceso se llevó a cabo mediante transformaciones controladas sobre los archivos MIDI originales, preservando su estructura rítmica esencial mientras se introducían variaciones en la asignación de instrumentos.

Inicialmente, se definió un diccionario de mapeo que asociaba los valores de *pitch* MIDI con los instrumentos fundamentales de una batería jazzística moderna: tarola (*snare drum*), bombo (*bass drum*), hi-hat, ride y dos toms. Esta selección se basó en su predominancia dentro del género y su capacidad para generar texturas rítmicas complejas.

Utilizando la librería **miditok**, cada uno de los 200 solos originales fue procesado para generar 10 variaciones distintas. En cada iteración, se seleccionó aleatoriamente una nota dentro del solo y se modificó su *pitch* dentro de un rango de 3 a 10 semitonos, asegurando que la nueva asignación correspondiera a un instrumento válido dentro del diccionario establecido. Este enfoque permitió alterar el color tímbrico de ciertos pasajes sin afectar la métrica, la dinámica ni la secuencia rítmica original, manteniendo así la coherencia estilística del jazz; el diagrama de flujo de la síntesis de archivos MIDI se puede observar en la figura 4.1.

Las versiones generadas se almacenaron conservando el nombre del archivo MIDI base, añadiendo un sufijo numérico para identificar cada variante, ejemplo: *A love supreme_Solo_1.mid*,

A love supreme_Solo_2.mid, etcétera. Esta estrategia no solo enriqueció la base de datos cuantitativamente, sino que también facilitó el entrenamiento de los distintos modelos al exponerlos a un espectro más amplio de combinaciones instrumentales dentro de un mismo contexto rítmico y melódico. La base de datos sintética resultante está disponible en el repositorio del proyecto: `https://github.com/rodrums21/DrumTransformer-`.

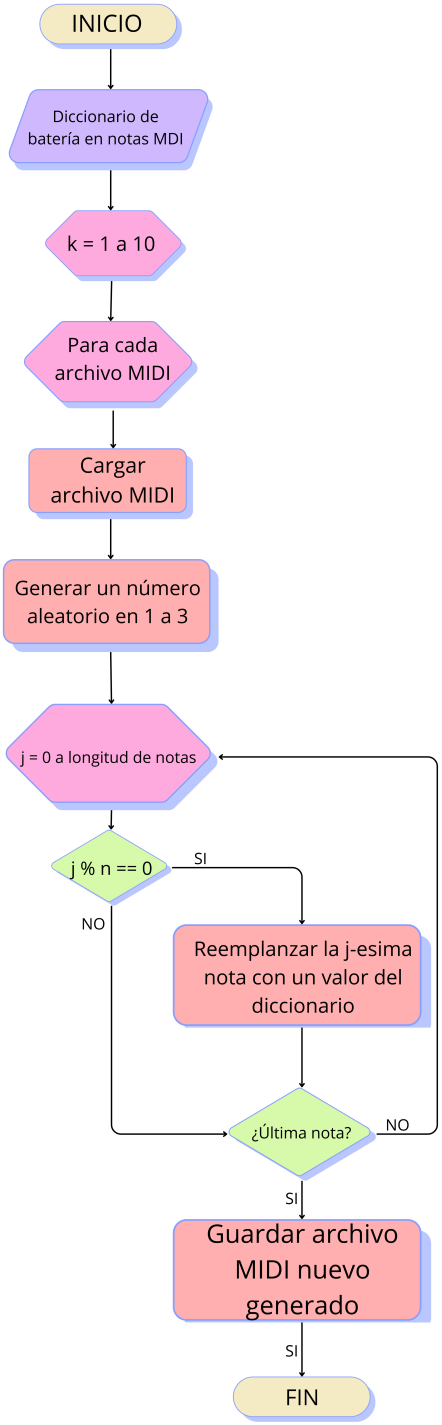


Figura 4.1. Diagrama de Flujo para la Síntesis de Archivos MIDI.

4.1.3. Preparación de las muestras para el entrenamiento del modelo

En el presente proyecto se implementó una variante simplificada de la arquitectura Transformer GPT-2, conocida como DistilGPT-2. La elección de este modelo se fundamentó en sus características técnicas, destacándose su reducción paramétrica a aproximadamente 82 millones de parámetros, en comparación con los 124 millones del modelo GPT-2 estándar [24]. Esta compresión del modelo resulta en una arquitectura sustancialmente más ligera, lo que facilita su implementación computacional y reduce significativamente los requisitos de recursos, manteniendo un rendimiento competitivo en las tareas de procesamiento de lenguaje natural para las cuales fue diseñado.

Antes de comenzar con el procesamiento de los archivos para generar las muestras, se llevaron a cabo dos acciones: en primer lugar, se depuró la base de datos, eliminando todos los solos que no se adecuaran a una estructura de 4/4. Posteriormente, se realizó una normalización que consistió en ajustar los tempos de cada solo a 120 bpm, teniendo un total de 198 solos útiles para el proceso de entrenamiento.

Como se muestra en la figura 4.2, el proceso de tokenización inició con la creación de un diccionario que vincula cada pitch de nota MIDI con la abreviatura correspondiente de los instrumentos de percusión. Por ejemplo, el número 36 está asociado con el bombo o kick drum en inglés, cuya abreviatura asignada fue "kck". La tabla 6.1 presenta una lista detallada que muestra cómo se asignaron las abreviaturas a cada nota.

Para desarrollar el token más pequeño para el entrenamiento del modelo, se implementó una función que utiliza la librería pretty_MIDI para acceder secuencialmente a cada instrumento y sus notas, como ilustra el flujo en la figura 4.2. El pitch se representó por un número entero y la duración por un número flotante estandarizado a dos decimales. Cada pitch se reemplazó por su abreviación correspondiente según el diccionario, concatenándola con la duración para formar tokens como `"snr,0.25"`.

Una vez establecida la función de tokenización, se procesaron sistemáticamente los archivos MIDI mediante su conversión a secuencias de tokens. Cada secuencia fue segmentada en ventanas de 500 y 1000 tokens, utilizando el carácter "|" como delimitador de fin de secuencia. La selección de estas longitudes se fundamentó en el análisis estadístico del corpus: los 1000 tokens corresponden aproximadamente al número promedio de eventos musicales contenidos en un compás estándar de jazz, mientras que la ventana de 500 tokens permitió duplicar el volumen de muestras manteniendo la integridad estructural de los patrones rítmicos.

Las secuencias tokenizadas se almacenaron en archivos de texto estructurados, optimizados para su compatibilidad con la librería **datasets**. El conjunto de datos resultante se dividió en un 70 % para entrenamiento y 30 % para validación, garantizando la representatividad de ambos subsets. Este proceso completo de preprocesamiento se documenta esquemáticamente en la Figura 4.2.

4.1.4. Entrenamiento del modelo

Para el entrenamiento del modelo DistilGPT2 fue empleado su tokenizador estándar. Tanto el modelo como el tokenizador fueron cargados de la librería Transformers de Hugging Face. Se configuró el modelo en modo de entrenamiento y se ajustó el tokenizador para utilizar el token de fin de secuencia como token de relleno, asegurando la consistencia en el manejo de secuencias de diferentes longitudes las cuales están predeterminadas por el tokenizador propio del modelo.

Las melodías se cargaron desde el archivo de texto previamente guardado, donde cada línea representaba una secuencia musical. Se utilizó el tokenizador de DistilGPT-2 para procesar los datos, aplicando padding y truncamiento para uniformizar la longitud. Las secuencias tokenizadas se almacenaron en un Dataset de Hugging Face, asignando los mismos identificadores de

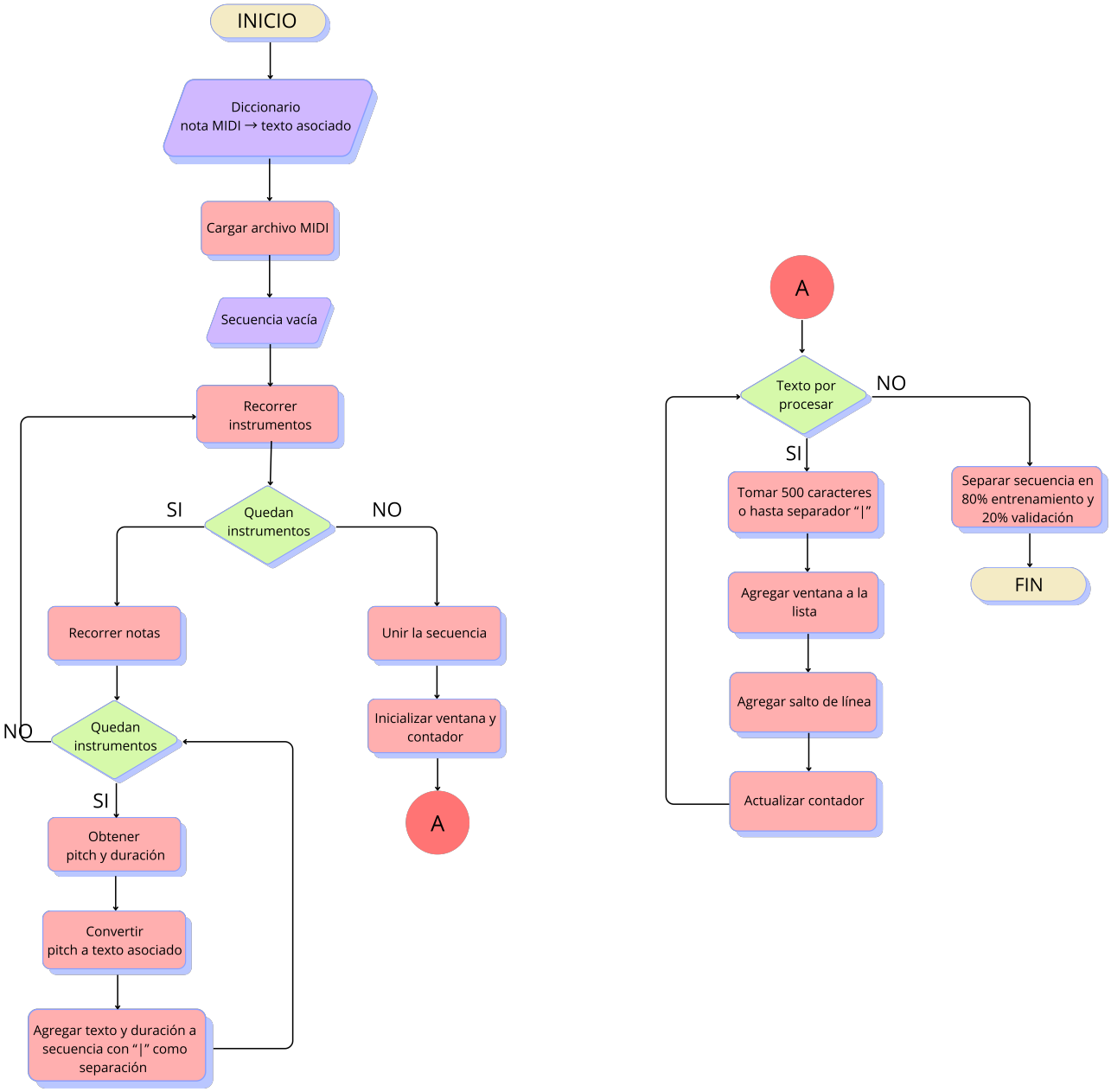


Figura 4.2. Tokenización de solos de batería jazz: conversión MIDI a secuencias textuales y creación de ventanas para DistilGPT2.

entrada como etiquetas para el entrenamiento auto regresivo.

Se realizó un fine-tunning del modelo, resultando en tres configuraciones con un tamaño de token de 500 y otras tres con un tamaño de token de 1000. La tasa de aprendizaje se estableció constante en 5×10^{-5}. Además, tanto el tamaño del lote como el valor del gradiente se mantuvieron constantes en 4. Se modificó el número de épocas para crear estas seis versiones, guardando modelos a las 6, 9 y 12 épocas de cada configuración para dar como resultado seis versiones distintas. Tanto el modelo como su tokenizador de cada versión fueron almacenados para usos futuros en generación de secuencias utilizando el método `save_pretrained()` de la biblioteca Transformers, y los archivos correspondientes se guardaron en una ruta designada para cada versión del modelo.

4.1.5. Generación de secuencias

Para la fase de generación, inicialmente se cargaron el modelo entrenado y su tokenizador correspondiente desde la ubicación donde fueron almacenados previamente. Como se detalla en la Figura 4.3, este proceso incluyó una verificación en dos etapas: primero se confirmó la integridad de los archivos del modelo y luego se validó que los parámetros musicales base (compás 4/4 y tempo 120 BPM) estuvieran correctamente configurados.

La generación se realizó mediante un proceso autoregresivo donde cada nuevo token musical se predice secuencialmente basado en los tokens anteriores. Este mecanismo, representado en el flujo iterativo de la Figura 4.3, utilizó dos parámetros clave para controlar la creatividad: temperatura, con un valor de 0.75 y top-k de 50, dichos valores fueron seleccionados de manera experimental hasta encontrar secuencias que mantuvieran una coherencia y estructura musical del jazz. La temperatura ajusta la aleatoriedad en las predicciones; valores más bajos producen salidas más conservadoras, mientras que valores más altos fomentan la diversidad. El parámetro top-k limita la selección del siguiente token considerando solo las k opciones más probables en cada paso, equilibrando entre calidad y variedad musical.

Cada iteración del proceso generativo, ilustrado en el bucle de la Figura 4.3, comenzó con un patrón rítmico inicial y produjo secuencias que fueron seleccionadas para conservar únicamente los fragmentos con estructura válida de nota y duración. Las secuencias aprobadas se convirtieron en archivos MIDI completos con todos los metadatos requeridos, incluyendo el compás 4/4 y tempo 120 BPM especificados en el diagrama.

4.2. Validación del modelo tipo Transformer generando un conjunto de 90 solos de batería

La evaluación del modelo de generación de solos de batería de jazz se realizó mediante un enfoque cualitativo basado en el juicio de 8 bateristas expertos en el género del jazz. El objetivo principal fue valorar el estilo, originalidad y viabilidad técnica de los solos generados por las diferentes configuraciones del modelo entrenado, comparándolos con ejemplos humanos de referencia. Para ello, se diseñó un proceso sistemático que incluyó la selección de muestras, la creación de instrumentos de evaluación y la estandarización en la presentación de los estímulos auditivos.

4.2.1. Descripción de la Muestra

La muestra evaluada consistió en 40 solos de batería, divididos en dos grupos:

- Solos generados: 30 solos producidos por las seis versiones del modelo, con cinco ejemplos seleccionados aleatoriamente para cada configuración.
- Solos humanos: 10 solos extraídos aleatoriamente de la base de datos original de 200 solos transcritos.

4.2.2. Instrumentos de Evaluación

Se utilizó un cuestionario estructurado a través de Google Forms que consistía en cinco preguntas. Cada una de estas preguntas adoptó una escala Likert [43] de cinco niveles, que variaban desde *"Totalmente en desacuerdo"* hasta *"Totalmente de acuerdo"*. Todas las preguntas se elaboraron de manera que cada una de ellas evaluara una dimensión particular de los solos. Las preguntas formuladas y las dimensiones son las siguientes:

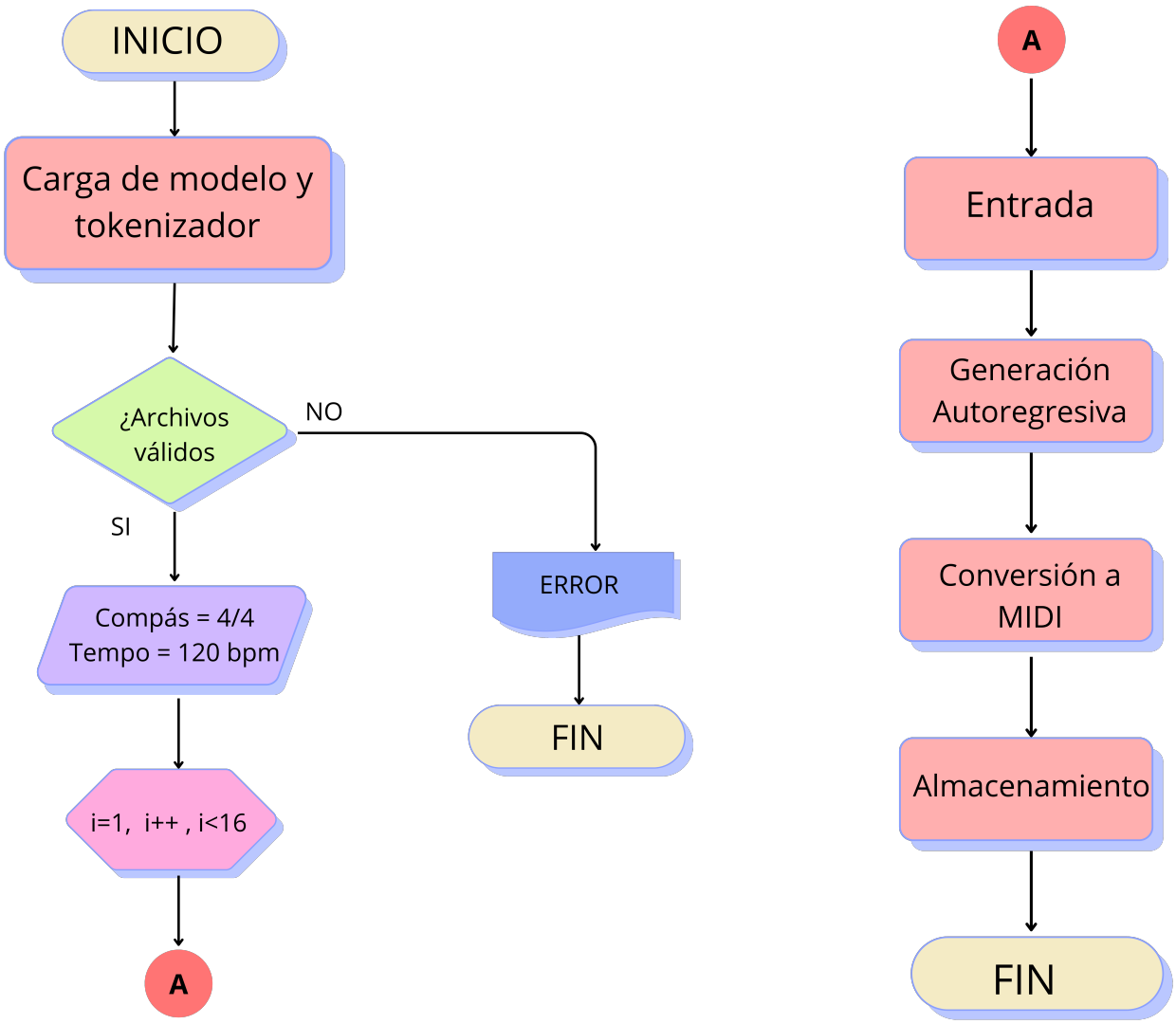


Figura 4.3. Diagrama de flujo para la generación de solos.

1. Estilo: ¿El ritmo del solo suena natural y apropiado para el jazz?
2. Coherencia estructural: ¿Los patrones del solo tienen lógica musical?
3. Originalidad: ¿El solo es repetitivo o genérico?
4. Ejecución: ¿Consideras que este solo puede ser ejecutado por una persona baterista profesional?
5. Percepción de composición: ¿Crees que este solo fue generado con alguna herramienta de Inteligencia Artificial?

4.2.3. Formato de Presentación de los Solos

Para garantizar una evaluación consistente, todos los solos se presentaron en formato WAV a 44.1 kHz, utilizando el sintetizador por defecto de MuseScore. Esta estandarización aseguró que la calidad de audio fuera uniforme en todas las muestras y que los evaluadores no se vieran influenciados por diferencias técnicas en la reproducción. Los archivos se organizaron y presentaron de manera aleatoria a los expertos, evitando así posibles sesgos en el orden de escucha.

Capítulo 5

Resultados

5.1. Originalidad y contribución en el contexto del estado del arte

Tras una revisión exhaustiva del estado del arte, se identificó que, si bien existen numerosos estudios centrados en el análisis y la generación de patrones rítmicos en jazz, la mayoría aborda la batería desde un rol secundario, como instrumento de acompañamiento o soporte armónico. No se encontraron investigaciones previas que exploraran de manera específica y protagonista los solos de batería dentro de un marco computacional, utilizando bases de datos especializadas y modelos de inteligencia artificial para su generación y evaluación. Esta brecha en la literatura resalta la originalidad del presente trabajo, que no solo propone un enfoque novedoso para el estudio de la improvisación percusiva en el jazz, sino que también establece un precedente en la aplicación de técnicas avanzadas, como los modelos Transformer, para emular y analizar la complejidad rítmica y expresiva característica de los solos de batería profesionales.

5.2. Base de datos de solos de batería de jazz

Como uno de los principales resultados, se desarrollaron dos bases de datos complementarias orientadas al análisis de patrones rítmicos en el jazz tradicional.

Los solos incluidos abarcan de 1935 al 2018, representativos de diferentes épocas y estilos dentro del género. Cada entrada fue verificada mediante un proceso de validación que garantizó la precisión de los metadatos y la integridad de los archivos MIDI. La base de datos permite consultas cruzadas por baterista, período o disco.

La primera base de datos recopila 200 solos de batería de jazz tradicional en formato MIDI, estructurados bajo seis variables clave: **baterista**, **canción**, **año de grabación**, **artista principal**, **disco** y **número de compases**. Esta colección, disponible públicamente en `https://github.com/rodrums21/DrumTransformer-`, constituye un recurso estandarizado para el análisis computacional de patrones rítmicos en el jazz tradicional.

El análisis de la distribución temporal de los 200 solos que componen a esta base de datos revela importantes variaciones en la representación histórica del jazz, como se muestra en la tabla 5.1. La década de 1950 destaca como el período mejor representado, con 113 solos (56.5 % del total). Le sigue la década de 1960 con 58 solos (29 %). Las décadas formativas (1930-1940) presentan una representación más modesta, con 3 solos (1.5 %) y 9 solos (4.5 %) respectivamente. La representación disminuye notablemente en décadas posteriores: 7 solos (3.5 %) en los 70, 3 (1.5 %) en los 80, 6 (3 %) en los 90, 4 (2 %) en los 2000s y solo 7 (3.5 %) en la década de 2010.

Esta distribución desigual refleja la coincidencia de dos factores principales. En primer lugar, la mayor disponibilidad de transcripciones correspondientes a períodos tradicionalmente

Década	Número de Solos	Porcentaje (%)
1930	3	1.5
1940	9	4.5
1950	113	56.5
1960	58	29
1970	7	3.5
1980	3	1.5
1990	6	3
2000	4	2
2010	7	3.5
Total	**200**	**100.00**

Cuadro 5.1. Distribución temporal de solos en la base de datos: cantidad y porcentaje por década (1930-2010).

categorizados como clásicos dentro de la historia del jazz. En segundo término, un enfoque curatorial que priorizó estilos específicos, en particular el bebop y el hardbop, en detrimento de otras vertientes y desarrollos históricos del género.

Es relevante destacar que el período comprendido entre las décadas de 1940 y 1960 corresponde a la consolidación de la batería como instrumento con una función solista predominante, lo cual explica parcialmente su sobrerrepresentación en el corpus. La delimitación temporal del conjunto de datos, que excluye composiciones posteriores al año 2018, constituyó una decisión metodológica. Si bien se reconoce la existencia de material reciente de potencial utilidad para el entrenamiento de modelos, esta selección acotada permite un enfoque controlado en las características técnicas de los períodos históricos mejor documentados.

A partir de esta colección, se generó una segunda base de datos sintética, disponible en el mismo repositorio . Esta versión ampliada conserva las mismas variables, pero introduce modificaciones algorítmicas en las pistas MIDI, reemplazando aleatoriamente el instrumento original mientras se mantiene la estructura rítmica base. El conjunto sintético fue escalado a 2000 entradas, proporcionando un recurso ampliado para entrenamiento de modelos computacionales y estudios de generalización en análisis musical.

Estos recursos han sido desarrollados para su aplicación en investigaciones dentro de campos como la inteligencia artificial aplicada a la música, la musicología u otras ramas. Proporciona un corpus accesible y reproducible, destinado a facilitar futuros estudios en estas disciplinas. La disponibilidad pública de estos recursos promueve la reproducibilidad y el avance en el campo.

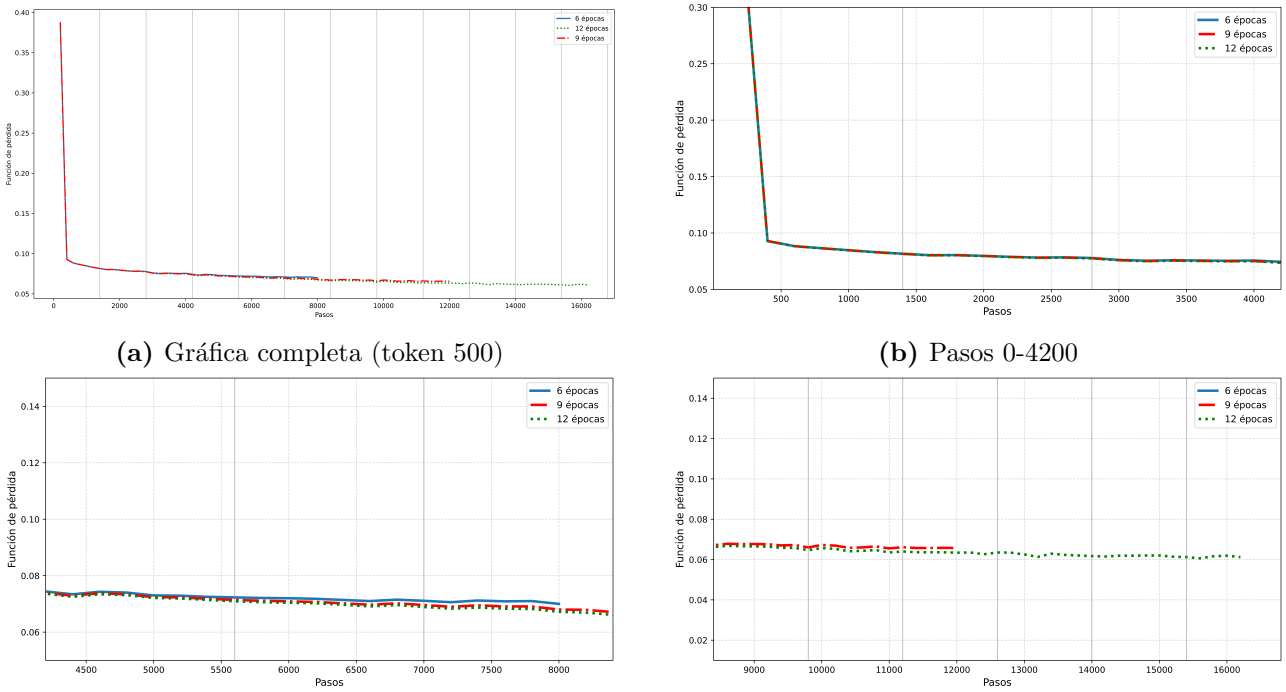
5.3. Modelo de redes neuronales tipo Transformer para la generación de solos de batería

5.3.1. Entrenamiento del modelo

Antes de proceder al análisis de las curvas de la función de pérdida, es necesario establecer la relación entre épocas y pasos para cada configuración. En los modelos entrenados con un tamaño de token de 500, una época corresponde a 1400 pasos, mientras que para el modelo con un tamaño de token de 1000, una época equivale a 700 pasos. Esta diferencia en la relación épocas-pasos es fundamental para interpretar correctamente la dinámica de convergencia durante el entrenamiento.

El análisis comparativo de las funciones de pérdida revela patrones diferenciados entre los

modelos entrenados con distintos tamaños de token. Como se observa en la figura 5.1, correspondiente al token de tamaño 500, las tres configuraciones muestran una marcada reducción inicial durante las primeras épocas. La configuración de 6 épocas presenta un descenso pronunciado desde 0.40 hasta 0.25 en la primera época, seguido de una moderación progresiva que la lleva a estabilizarse alrededor de 0.15.



(a) Gráfica completa (token 500)

(b) Pasos 0-4200

(c) Pasos 4200-8400

(d) Pasos 8400-16800

Figura 5.1. Análisis de función de pérdida para token tamaño 500. Se muestran la evolución completa y tres intervalos significativos del proceso de entrenamiento.

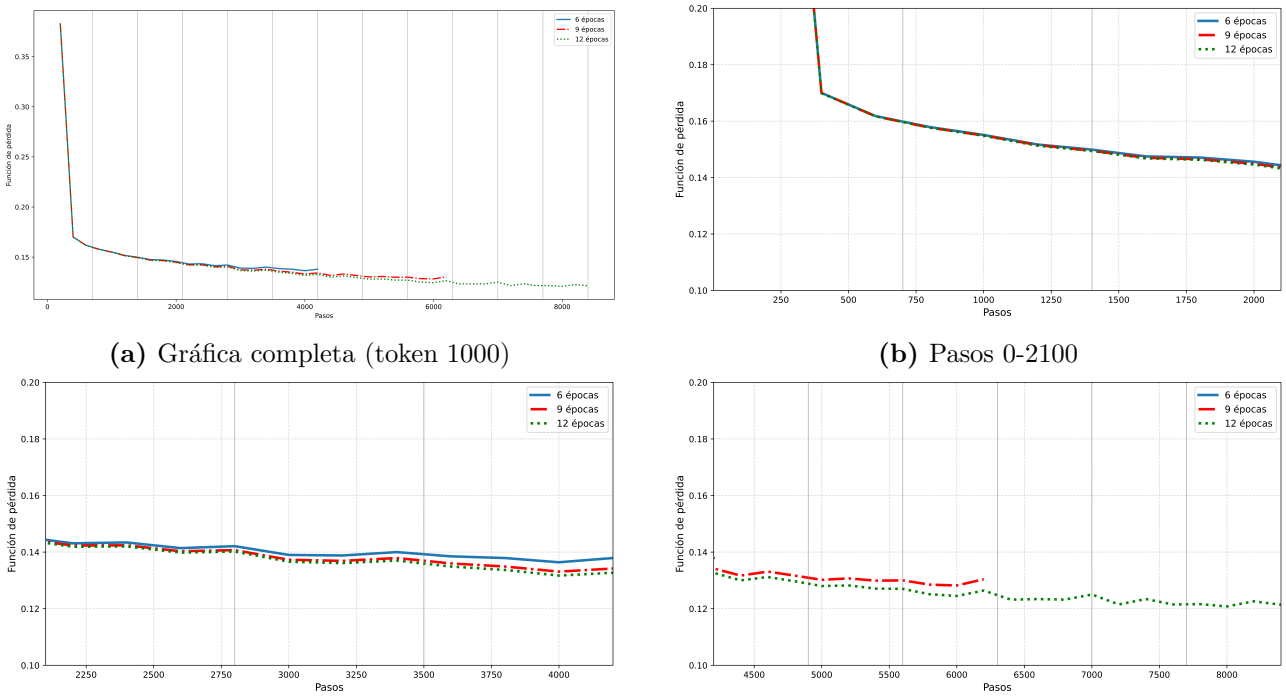
Esta tendencia se amplifica en las configuraciones de mayor duración, donde la de 9 épocas alcanza 0.12 y la de 12 épocas converge cerca de 0.10, como puede apreciarse en la figura 5.1d. La tabla 5.2 resume los valores clave obtenidos en cada configuración.

En contraste, la figura 5.2 muestra un comportamiento distinto para el token de 1000. Las curvas exhiben una reducción más uniforme a lo largo del entrenamiento, sin las fases marcadas que caracterizan al token menor. La configuración de 6 épocas desciende constantemente desde 0.35 hasta 0.25, mientras que las de 9 y 12 épocas mantienen pendientes sostenidas, alcanzando 0.20 y 0.18, respectivamente.

Configuración	Token 500	Token 1000
6 épocas	0.15	0.25
9 épocas	0.12	0.20
12 épocas	0.10	0.18

Cuadro 5.2. Resumen de valores de función de pérdida por configuración

El detalle de la figura 5.2c muestra cómo estas configuraciones no presentan puntos de inflexión abruptos, sugiriendo un proceso de aprendizaje más gradual. La comparación entre ambos conjuntos de gráficas evidencia dinámicas de aprendizaje fundamentalmente diferentes. Mientras el token 500 (figura 5.1b) muestra una fase inicial de rápida convergencia que abarca aproximadamente el 65 % de la reducción total, seguida de un ajuste fino progresivo, el token 1000 mantiene una mejora constante sin etapas claramente delimitadas.



(a) Gráfica completa (token 1000)

(b) Pasos 0-2100

(c) Pasos 2100-4200

(d) Pasos 4200-8400

Figura 5.2. Análisis de función de pérdida para token tamaño 1000. Se muestran la evolución completa y tres intervalos significativos del proceso de entrenamiento.

Esta variación en los comportamientos podría atribuirse a la mayor complejidad de representación requerida por el token de 1000, que demandaría un proceso de entrenamiento más prolongado para alcanzar niveles equivalentes de optimización. Los detalles mostrados en las figuras 5.1d y 5.2d muestran claramente esta diferencia, donde el primer caso muestra estabilización, mientras que el segundo mantiene pendiente de descenso.

El modelo DistilGPT-2 fue configurado para generar solos de batería de jazz mediante dos longitudes de entrada y tres duraciones de entrenamiento, produciendo 90 muestras en total. Cada combinación generó resultados musicalmente diversos, desde patrones rítmicos concisos hasta desarrollos más extensos.

5.4. Validación de los modelos tipo Transformer generando un conjunto de 90 solos de batería

Durante la evaluación realizada sobre la muestra de solos de batería, tanto humanos como generados por IA, se analizó la capacidad de bateristas profesionales para distinguir entre ambos orígenes. Los resultados, representados en la matriz de confusión de la figura 5.3, mostraron que el 65.9 % de los solos humanos fueron identificados correctamente como reales, mientras que el 62.1 % de los solos generados por el modelo Transformer fueron reconocidos acertadamente como IA. No obstante, un 34.1 % de los solos reales fueron erróneamente atribuidos a la inteligencia artificial, y un 37.9 % de los solos generados fueron percibidos como ejecutados por humanos, lo que evidencia cierto grado de confusión en la identificación y sugiere que los solos generados poseen características que en ocasiones se aproximan a la expresividad humana.

El proceso de evaluación de los modelos de Transformer se fundamentó en la capacidad de bateristas profesionales para discernir entre solos compuestos por humanos y aquellos generados automáticamente por los diferentes modelos del Transformer. Los resultados cuantitativos,

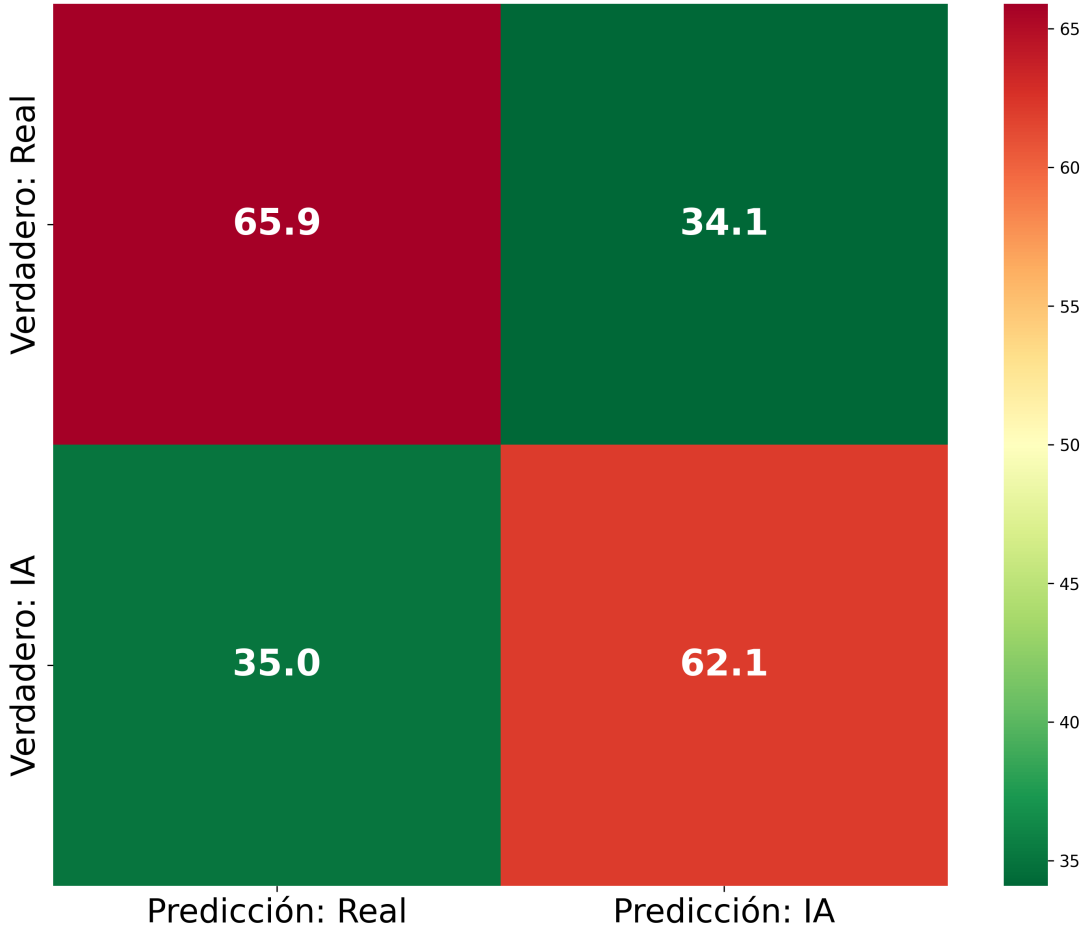


Figura 5.3. Matriz de confusión que muestra el porcentaje de identificación correcta y errónea de solos reales y generados por IA.

sintetizados en la matriz de calor de la Figura 5.4, revelan patrones de rendimiento respecto a la eficacia de cada configuración de hiperparámetros. Este análisis visual permite identificar de manera clara no solo el desempeño individual de cada modelo, sino también las interacciones y tendencias relacionadas entre el tamaño del token y las épocas de entrenamiento.

La configuración que demostró un desempeño óptimo y superior al resto fue la que empleó un tamaño de token de 500 y un ciclo de 6 épocas de entrenamiento, alcanzando un notable 50.0 % de efectividad. Este resultado indica que, en una muestra representativa, la mitad de las composiciones generadas por este modelo fueron percibidas por los expertos como obras creadas por bateristas profesionales.

En un segundo lugar en el rendimiento, claramente diferenciado en la matriz de calor por sus tonos intermedios, lo ocupa el modelo de 1000 tokens y 9 épocas, el cual registró un 40.0 % de eficacia. Muy de cerca, y formando un grupo de similar desempeño, se encuentran tres configuraciones: el modelo de 500 tokens con 12 épocas y el de 1000 tokens con 6 épocas, ambos con un 35.0 % de precisión, seguido por el modelo de 1000 tokens y 12 épocas, que obtuvo un 32.5 %. La cercanía entre estos porcentajes sugiere que estas combinaciones de parámetros tienen un rendimiento muy similar. Esto indica que, para estas configuraciones específicas, aumentar la complejidad del modelo o alargar el tiempo de entrenamiento no produce una mejora significativa en la calidad del resultado final.

Por último, y como se evidencia visualmente en el color más frío de su celda correspondiente en la matriz, la configuración con el desempeño más modesto fue la de 500 tokens y 9 épocas, la cual alcanzó solo un 30.0 % de identificación correcta. Este resultado va en contra de lo que se esperaría normalmente, pues el mismo tamaño de token de 500 que da el mejor resultado

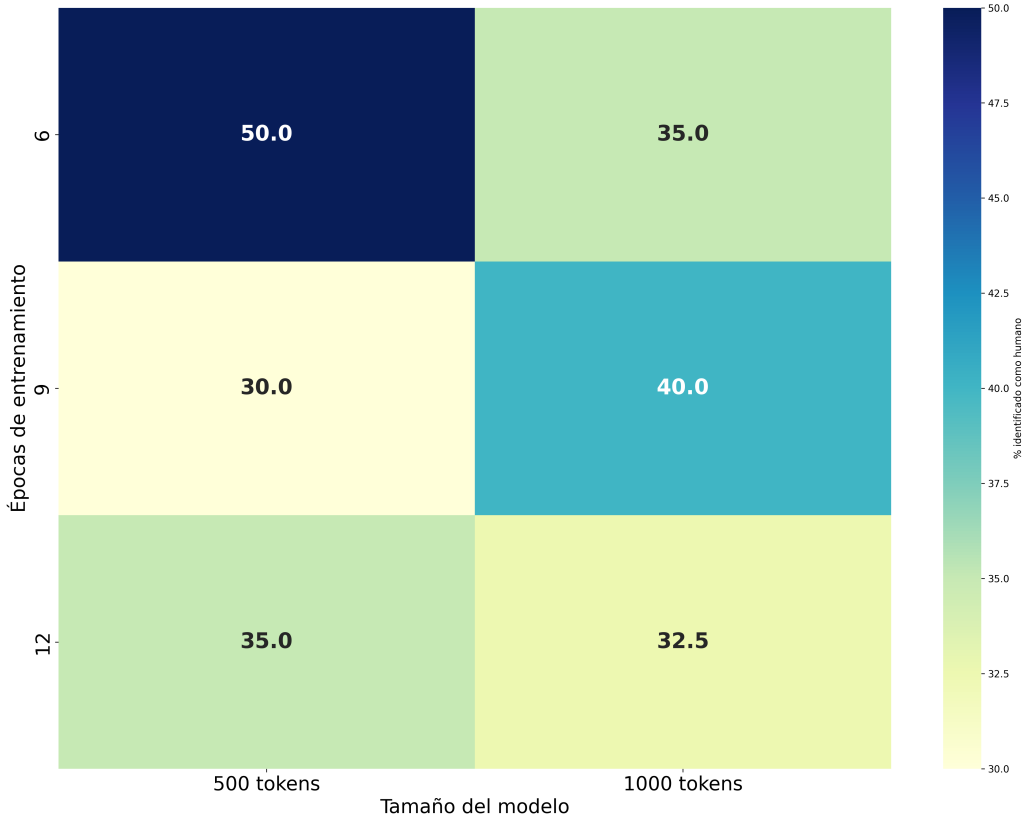


Figura 5.4. Matriz de calor del rendimiento de los modelos Transformer evaluados. La intensidad del color en cada celda representa el porcentaje de solos generados por cada modelo que fueron identificados como compuestos por un humano.

también produce el peor, solo por cambiar las épocas de entrenamiento. Esto fundamenta la idea de que la relación entre los ajustes del modelo es compleja.

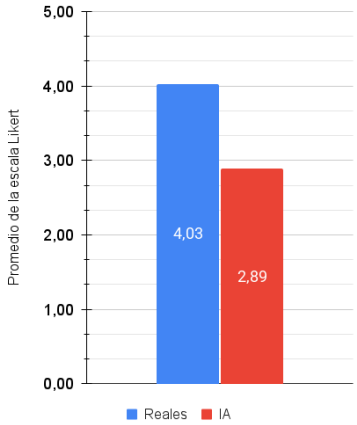
El análisis general de la matriz de calor permite llegar a una conclusión, la cual no muestra una tendencia clara que indique que un mayor tamaño de token o más épocas de entrenamiento produzcan mejores resultados. Por el contrario, la efectividad de cada modelo depende de una combinación específica y equilibrada de ambos parámetros. La configuración de 500 tokens con 6 épocas emerge claramente como la más adecuada para esta tarea en particular, demostrando un mejor desempeño que otras configuraciones que, aparentemente, eran más complejas o habían recibido un entrenamiento más prolongado. Este hallazgo resalta la importancia de realizar una búsqueda de la combinación ideal de los hiperparámetros al implementar modelos para la generación de música, ya que la optimización de estos parámetros resulta determinante para el éxito del modelo.

5.5. Evaluación cualitativa de los solos generados

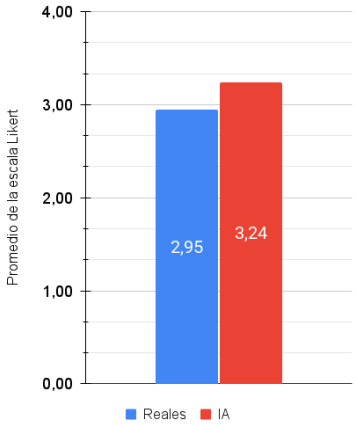
5.5.1. Comparación entre solos reales contra solos generados

En la metodología, se planteó la evaluación de 4 propiedades clave para cada solo generado o real, utilizando una escala Likert de 1 a 5. En la presentación de estos resultados se están utilizando los promedios obtenidos en la evaluación de estas características. Primero, se compararon los solos humanos con los solos artificiales, según la figura 5.5. En cuanto al **estilo**, los expertos dieron un promedio de 3.90 a los solos auténticos y 2.70 a los generados, como se muestra en la figura 5.5a. Para la **ejecución humana**, los solos reales obtuvieron un promedio de 4.54,

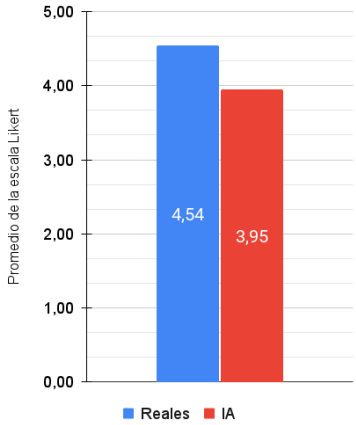
superando al 3.95 conseguido por los solos generados, según la figura 5.5c. En lo que respecta a la **coherencia estructural**, los solos auténticos alcanzaron un puntaje de 4.03, mientras que los generados registraron 2.89, como se muestra en la figura 5.5d. La única categoría donde los solos generados destacaron fue en **originalidad**, con una puntuación de 3.2 frente a 2.9 para los solos reales, de acuerdo con la figura 5.5b.



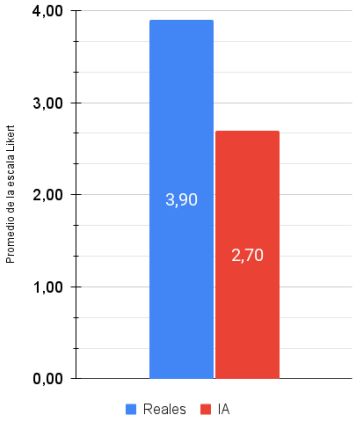
(a) Estilo.



(b) Originalidad.



(c) Ejecución Humana.



(d) Coherencia Estructural

Figura 5.5. Gráficas del análisis cualitativo entre los solos reales y los generados por los modelos entrenados del Transformer

Los hallazgos revelan un rendimiento contrastante entre los solos auténticos y los producidos por modelos Transformer, mostrando tanto las capacidades como las limitaciones actuales de estos sistemas. En particular, en las dimensiones de estilo con valores de 3.90 frente a 2.70 y coherencia estructural con 4.03 frente a 2.89, los solos humanos presentan una ventaja notable. La diferencia más significativa se encuentra en la coherencia estructural, con una disparidad de 1.14 puntos. Esta disparidad era esperable, ya que el entrenamiento de los modelos omite elementos expresivos fundamentales como la dinámica y la articulación en la interpretación instrumental, que son cruciales para lograr una ejecución que sea tanto orgánica como estilísticamente convincente.

Aunque en la ejecución realizada por humanos la diferencia se estrecha a 4.54 frente a 3.95, con una variación de 0.59 puntos, esto indica que los modelos son capaces de replicar de manera aceptable algunos aspectos técnicos del desempeño profesional. Además, en términos de originalidad, los solos generados superan ligeramente a los creados por humanos, con puntuaciones de 3.2 frente a 2.9, lo que sugiere que los sistemas tienen una notable aptitud para generar ideas novedosas, pese a una menor cohesión en la estructura.

Los resultados de este estudio revelan una dicotomía significativa: aunque los modelos muestran un notable potencial para crear material distintivo y adaptarse a las normas de estilo del jazz, todavía afrontan dificultades para imitar la naturalidad, fluidez y expresión que los músicos aportan. La relativa fortaleza de los modelos en originalidad se contrapone a sus deficiencias en el ámbito estructural y estilístico, lo que sugiere que, si bien la generación artificial tiene la capacidad de crear solos creativos, todavía no logra alcanzar la capacidad interpretativa propia de una ejecución humana. En su conjunto, los resultados indican un desarrollo en la síntesis de solos de batería de jazz, aunque destacan áreas críticas donde todavía es necesario mejorar, especialmente en lo referente a la integración de elementos expresivos y estructurales que son esenciales para lograr una interpretación humana auténtica.

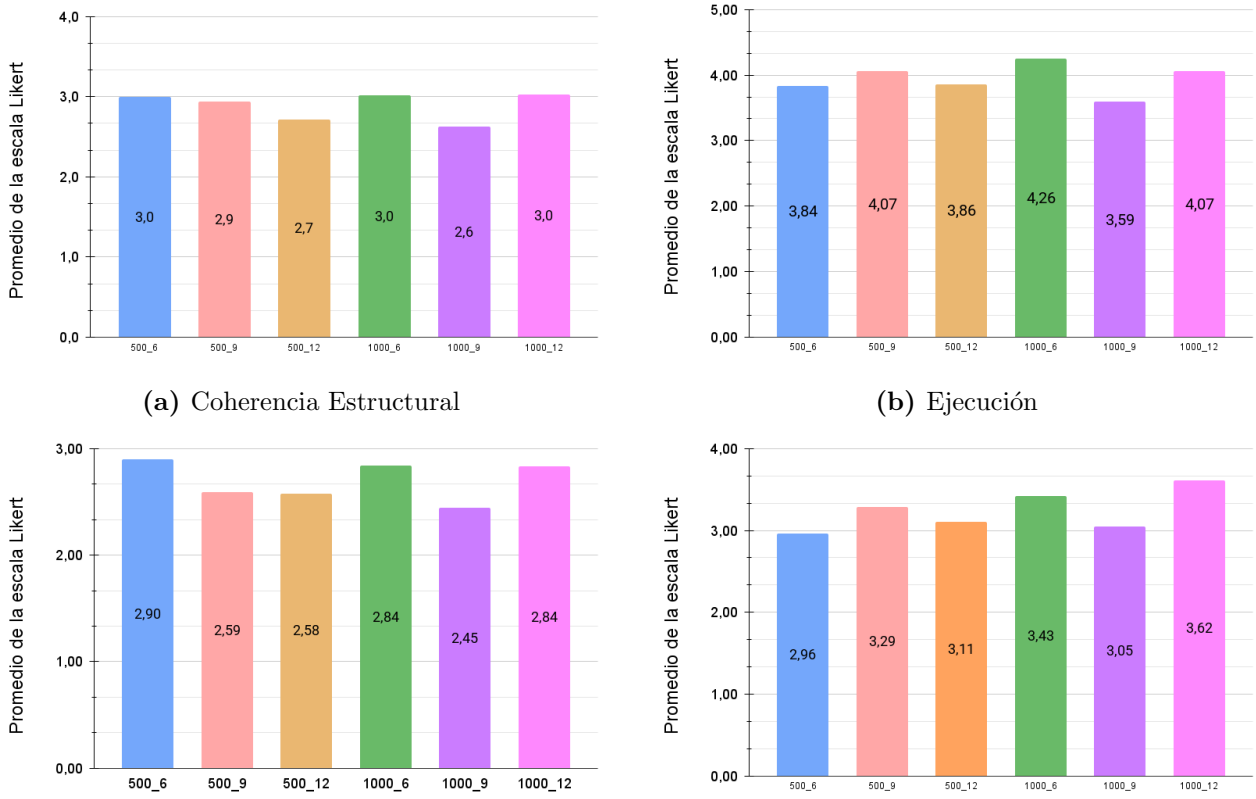
5.5.2. Evaluación cualitativa por cada modelo

El análisis de los modelos reveló diferencias significativas en su capacidad para generar solos de batería de jazz en función del tamaño del token y las épocas de entrenamiento , como se muestra en la figura 5.6. En la categoría de **ejecución humana** , figura 5.6b, el modelo con un tamaño de token de 1000 y 12 épocas de entrenamiento obtuvo la puntuación más alta, 4.26; seguido por el modelo de 500 tokens y 9 épocas con una puntuación de 4.07. En **coherencia estructural**, figura 5.6a, los modelos de 1000 tokens superaron consistentemente a los de 500 tokens, destacando especialmente aquellos entrenados durante 9 y 12 épocas.

Sin embargo, en categorías más subjetivas como **estilo** (figura 5.6c), ningún modelo superó los 2.90 puntos, aunque el de 1000 tokens y 12 épocas obtuvo el mejor desempeño. En cuanto a **originalidad**, figura 5.6d, el modelo de 1000 tokens y 9 épocas superó ligeramente al de 12 épocas, con puntuaciones de 3.62 y 3.43, respectivamente.

La investigación reveló una correlación entre los hiperparámetros del modelo, como el tamaño de los tokens y las épocas de entrenamiento, y su eficacia en la generación de solos de batería de jazz. En particular, el modelo configurado con 1000 tokens y 12 épocas evidenció el mejor desempeño general, sobresaliendo en cuanto a ejecución humana, con una puntuación de 4.26, y coherencia estructural. Este hallazgo sugiere que incrementar tanto la capacidad del tamaño de los tokens utilizados para entrenar el modelo como la duración del entrenamiento puede mejorar notablemente la naturalidad y la solidez estructural de los resultados. No obstante, el rendimiento casi comparable del modelo de 500 tokens con 9 épocas, que logró una puntuación de 4.07 en ejecución humana, sugiere la existencia de un punto óptimo. En este equilibrio, el tamaño del modelo y la duración del entrenamiento alcanzan un nivel donde aumentarlos más allá de cierto umbral podría no traducirse en mejoras significativas adicionales.

En comparación, los elementos más creativos, como el estilo y la originalidad, obtuvieron calificaciones significativamente más bajas; ningún modelo alcanzó más de 2.90 en estilo, lo que pone de manifiesto una limitación crucial en los enfoques actuales.Aunque el modelo que utilizó 1000 tokens durante 9 épocas obtuvo la mayor puntuación en cuanto a originalidad, con un 3.62, seguido por la misma configuración tras 12 épocas, que logró un 3.43, la mínima diferencia sugiere que la originalidad depende menos de un entrenamiento extenso y más de otros factores, posiblemente relacionados con la arquitectura del modelo o la variabilidad del conjunto de datos. Esta discrepancia entre el rendimiento en métricas técnicas, la ejecución humana y la coherencia, y los atributos artísticos como el estilo y la originalidad, resalta un desafío crucial para la investigación futura: cómo integrar criterios de creatividad y expresión personal en los sistemas de generación de música impulsados por inteligencia artificial.



(a) Coherencia Estructural

(b) Ejecución

(c) Estilo

(d) Originalidad

Figura 5.6. Evaluación cualitativa de los modelos por configuración (tokens/épocas) en: (a) Coherencia, (b) Ejecución, (c) Estilo y (d) Originalidad.

5.6. Limitaciones del estudio

El presente estudio, si bien innovador en su enfoque sobre la generación de solos de batería de jazz mediante modelos Transformer, presenta varias limitaciones técnicas y metodológicas que deben ser consideradas al interpretar los resultados. Estas limitaciones se agrupan en cuatro categorías principales: **limitaciones de los datos**, **limitaciones del modelo**, **limitaciones en la evaluación** y **limitaciones musicales**.

Limitaciones de los datos

La base de datos utilizada, aunque representativa de ciertos períodos y estilos del jazz, presenta una distribución desigual en cuanto a décadas, con una mayor concentración en las décadas de 1950 y 1960. Esto podría sesgar el modelo hacia patrones rítmicos característicos de esos períodos, limitando su capacidad para capturar innovaciones técnicas y estilísticas de épocas más recientes. Además, la generación de la base de datos sintética, aunque amplió el conjunto de entrenamiento, se basó en modificaciones algorítmicas que podrían no reflejar completamente la diversidad y complejidad de los solos humanos.

Limitaciones del modelo

La elección de DistilGPT-2, si bien eficiente computacionalmente, impone restricciones en la capacidad del modelo para capturar matices expresivos cruciales en la ejecución jazzística, como dinámicas, articulaciones y ghost notes. La tokenización empleada, con ventanas de 500 y 1000 tokens, tampoco logra capturar adecuadamente las macroestructuras musicales, lo que se

refleja en las bajas puntuaciones de coherencia estructural. Además, el entrenamiento en Google Colab Pro introdujo limitaciones técnicas, como la desconexión automática tras 12 horas de inactividad y restricciones en la capacidad de cómputo, lo que pudo afectar la extensión y profundidad de los experimentos.

Limitaciones en la evaluación

El protocolo de evaluación, aunque riguroso, presenta ciertas limitaciones metodológicas. La muestra de evaluación, compuesta por 40 solos, podría no ser suficientemente representativa de la diversidad estilística del jazz. La escala Likert utilizada introduce subjetividad en la medición de constructos como "originalidad" o "estilo". Además, la evaluación se realizó exclusivamente mediante archivos WAV generados con el sintetizador por defecto de MuseScore, lo que podría haber afectado la percepción de autenticidad debido a limitaciones en la calidad sonora.

Limitaciones musicales

Los resultados revelan brechas significativas entre los solos generados y los humanos. Los modelos mostraron dificultades para emular la intencionalidad artística y la sensibilidad estilística característica de los bateristas profesionales. La métrica de originalidad, aunque favorable para los solos generados, podría estar reflejando patrones inusuales más que verdaderas innovaciones musicales. Estas limitaciones destacan la complejidad de replicar la creatividad y la expresión humana en sistemas de inteligencia artificial.

5.7. Trabajo a futuro

Inicialmente, el análisis de la base de datos indica la necesidad de expandir y equilibrar el corpus actual, que tiene una sobrerrepresentación de las décadas de 1950 y 1960. Se recomienda enriquecerlo con transcripciones detalladas de bateristas pioneros de las décadas de 1930 y 1940. Asimismo, se debería incorporar metadatos avanzados sobre dinámicas y articulaciones utilizando herramientas como MuseScore. Adicionalmente, se sugiere el desarrollo de algoritmos de aumento de datos que preserven las características estilísticas del jazz tradicional, de un modo similar a la síntesis descrita en la metodología, pero con un control más preciso sobre las características musicales típicas del jazz. También se propone el entrenamiento del modelo con únicamente los solos correspondientes a las décadas de 1950 y 1960 y observar si el rendimiento del mismo difiere a los resultados presentados en esta investigación.

En segundo lugar, el modelo Transformer revela limitaciones importantes en términos de la coherencia estructural y la ejecución por parte de humanos, a pesar de funcionar bien en originalidad. Para solucionar estos problemas, se sugiere crear arquitecturas híbridas que integren la capacidad generativa de los Transformers con modelos basados en conocimiento musical explícito. Además, se deberían implementar esquemas de tokenización especializados que hagan una distinción clara entre los patrones de acompañamiento y los pasajes solistas, junto con mecanismos de atención mejorados para captar las relaciones a largo plazo dentro de estructuras musicales típicas.

El tercer enfoque busca la optimización de los protocolos de evaluación. Los datos cualitativos indican que únicamente el 37.9 % de los solos generados fueron evaluados como profesionales, mientras que el 65 % de los solos tocados por humanos fueron identificados correctamente. Para disminuir esta diferencia, se propone elaborar evaluaciones más sólidas con la participación de bateristas especializados en jazz tradicional, además de desarrollar y aplicar métricas computacionales de estilo como el análisis de patrones rítmicos, la distribución de acentos, el swing ratio, y estudiar también la presentación de los solos. Sería beneficioso considerar el uso de otro

motor para la generación de los audios y enfatizar en la humanización de la interpretación de los solos mediante herramientas que ya están disponibles.

Tomando en cuenta las fortalezas evidenciadas en la creación de material original y las recomendaciones para aplicaciones pedagógicas delineadas en la metodología, se sugiere el desarrollo de materiales educativos interactivos que representen la evolución histórica de los solos de batería. Además, se plantea la creación de herramientas de apoyo a la improvisación dirigidas a estudiantes, que integren la generación automática con retroalimentación inmediata en tiempo real. También se propone implementar sistemas de co-creación entre humanos e inteligencia artificial para la composición musical, permitiendo al modelo responder coherentemente a las ideas rítmicas planteadas por el usuario, asegurando que el resultado sea estilísticamente consistente.

Capítulo 6

Conclusiones

El presente trabajo abordó la generación automática de solos de batería de jazz mediante modelos Transformer, un campo poco explorado en el estado del arte, que tradicionalmente ha relegado la batería a un rol secundario. Los resultados obtenidos demuestran avances significativos en la capacidad de estos modelos para emular patrones rítmicos característicos del jazz, aunque también revelan limitaciones en aspectos como la coherencia estructural y la expresividad musical. A continuación, se sintetizan los hallazgos más relevantes, sus implicaciones y las perspectivas futuras derivadas de esta investigación.

Uno de los principales aportes de este estudio fue la creación de dos bases de datos especializadas: una compuesta por 200 solos de batería en formato MIDI, representativos de distintas épocas y estilos del jazz, y otra sintética ampliada a 2000 entradas mediante modificaciones algorítmicas. Estas bases de datos, disponibles públicamente, constituyen un recurso valioso para futuras investigaciones en generación musical y análisis computacional. Sin embargo, el análisis de distribución temporal evidenció un desequilibrio notable, con 56.5 % de los solos correspondientes a la década de 1950 y 29 % a la de 1960, mientras que décadas como los años 30, 80 y 2000 presentaron una representación inferior al 5 %. Esta disparidad podría sesgar los modelos hacia estilos específicos, limitando su capacidad para capturar innovaciones técnicas y estilísticas de períodos menos representados.

En cuanto al modelo Transformer implementado, los resultados mostraron que la configuración con un tamaño de token de 500 y 6 épocas de entrenamiento alcanzó el mejor rendimiento, con un 50 % de eficacia en la generación de solos clasificados como auténticos por los evaluadores. En contraste, el modelo con 1000 tokens y 12 épocas obtuvo un 40 % de eficacia, mientras que otras combinaciones, como 500 tokens y 9 épocas, mostraron un desempeño inferior, con solo 35 %. Estos resultados sugieren que un mayor tamaño de token y épocas adicionales de entrenamiento no siempre garantizan mejoras significativas, e incluso podrían generar sobreajuste o estancamiento en la reducción de la función de pérdida. Por otro lado, el modelo con 1000 tokens y 12 épocas destacó en métricas cualitativas como ejecución humana y coherencia estructural, alcanzando puntuaciones de 4.26/5 y 3.62/5 respectivamente, lo que indica que configuraciones más complejas pueden favorecer la captura de patrones rítmicos sofisticados.

La evaluación cualitativa realizada por bateristas expertos reveló diferencias notables entre los solos generados y los auténticos. En términos de estilo, los solos humanos obtuvieron un promedio de 3.90, frente a 2.70 de los generados, mientras que en coherencia estructural la diferencia fue de 4.03 contra 2.89. Estos datos reflejan las dificultades de los modelos para emular la fluidez y lógica musical de las interpretaciones humanas. Sin embargo, en originalidad los solos generados superaron ligeramente a los humanos, con 3.2/5 frente a 2.9/5, lo que sugiere que la inteligencia artificial puede aportar ideas novedosas, aunque sin la cohesión estilística característica del jazz tradicional. La ejecución humana también mostró una brecha importante, con 4.54/5 para solos reales y 3.95/5 para los generados, indicando que aspectos como la

dinámica, la articulación y la expresividad siguen siendo desafíos críticos.

Entre las limitaciones identificadas destacan las restricciones del modelo DistilGPT-2, que, si bien es eficiente computacionalmente, carece de capacidad para capturar matices expresivos clave en la improvisación jazzística, como ghost notes o variaciones de velocidad. Además, la tokenización empleada, con ventanas de 500 y 1000 tokens, no logró representar adecuadamente las macroestructuras musicales, lo que se tradujo en bajas puntuaciones de coherencia estructural. Otra limitación fue el protocolo de evaluación, basado en archivos WAV generados con el sintetizador por defecto de MuseScore, lo que pudo afectar la percepción de autenticidad debido a limitaciones en la calidad sonora.

Las perspectivas de trabajo futuro se centran en tres áreas principales. En primer lugar, se recomienda expandir y equilibrar la base de datos, incorporando más solos de décadas poco representadas y añadiendo metadatos avanzados sobre dinámicas y articulaciones. En segundo lugar, se sugiere explorar arquitecturas híbridas que combinen Transformers con modelos basados en conocimiento musical explícito, así como esquemas de tokenización especializados. Finalmente, se propone optimizar los protocolos de evaluación, integrando métricas computacionales de estilo, como el análisis de patrones rítmicos y el swing ratio, y utilizando sintetizadores más avanzados para mejorar la calidad sonora de las muestras generadas.

En el ámbito práctico, los resultados indican que los solos generados podrían ser útiles como material pedagógico para estudiantes de batería, proporcionando ejemplos de patrones rítmicos y estructuras de solos. Asimismo, su alto puntaje en originalidad los convierte en herramientas prometedoras para la composición asistida por computadora, aunque aún distan de igualar la creatividad y sensibilidad de un músico experto.

En conclusión, este estudio demuestra que los modelos Transformer son capaces de generar solos de batería de jazz con un grado notable de originalidad y viabilidad técnica, aunque presentan limitaciones en estilo y coherencia estructural. Los avances logrados representan un paso importante hacia la integración de la inteligencia artificial en la creación musical, pero también subrayan la complejidad de emular la expresividad humana. Futuras investigaciones deberán abordar estas limitaciones mediante mejoras en los datos, las arquitecturas y los métodos de evaluación, con el objetivo de acercarse cada vez más a la riqueza artística del jazz interpretado por profesionales.

Anexo 1

Pith	Instrumento	Abreviatura
35	Bombo 2	bdr
36	Bombo 1	kck
37	Baqueta	stk
38	Tarola	snr
39	Palmada	clp
40	Tarola eléctrica	esn
41	Tom bajo	ltm
42	Hi-Hat cerrado	chh
43	Tom de piso	ftm
44	Hi-Hat de pedal	phh
45	Tom medio	mtm
46	Hi-Hat abierto	ohh
47	Tom medio-bajo	lmt
48	Tom medio-alto	hmt
49	Platillo crash 1	cym
50	Tom alto	htm
51	Platillo ride 1	ryd
52	Platillo chino	chn
53	Campana de ride	rbl
54	Pandereta	tam
55	Platillo splash	spl
56	Cencerro	cbl
57	Platillo crash 2	cy2
58	Vibráfono	vib
59	Platillo ride 2	ry2
60	Bongo alto	bng
61	Bongo bajo	bgl
62	Conga alta muteada	mhc
63	Conga alta abierta	ohc
64	Conga baja	lco
65	Timbale alto	htb
66	Timbale bajo	ltb
67	Agogó alto	hag
68	Agogó bajo	lag
69	Cábasa	cab
70	Maracas	mar
71	Silbato corto	swh
72	Silbato largo	lwh

73	Güiro corto	sgu
74	Güiro largo	lgu
75	Claves	clv
76	Bloque de madera alto	hwb
77	Bloque de madera bajo	lwb
78	Cuica muteada	mcu
79	Cuica abierta	ocu
80	Triángulo muteado	mtr
81	Triángulo abierto	otr

Cuadro 6.1. Diccionario Pitch/Abreviatura

Bibliografía

[1] T. Shultz, "A history of jazz drumming," *Percussive Notes*, vol. 37, no. 2, pp. 105–132, 1999.

[2] M. Kaliakatsos-Papakostas, A. Floros, and M. N. Vrahatis, "Chapter 13 - artificial intelligence methods for music generation: a review and future perspectives," in *Nature-Inspired Computation and Swarm Intelligence* (X.-S. Yang, ed.), pp. 217–245, Academic Press, 2020.

[3] L. Moysis, L. A. Iliadis, S. Sotiroudis, A. Boursianis, M. Papadopoulou, K.-I. Kokkinidis, C. Volos, P. Sarigiannidis, S. Nikolaidis, and S. Goudos, "Music deep learning: Deep learning methods for music signal processing - a review of the state-of-the-art," *IEEE Access*, vol. PP, pp. 1–1, 01 2023.

[4] S. Yang, "The illusion of "authenticity": Ethical dilemmas and aesthetic imagination in pop music creation in the age of ai," *Journal of Contemporary Art Criticism*, vol. 1, p. 28–31, May 2025.

[5] J. Barnett, "The ethical implications of generative audio models: A systematic literature review," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, (New York, NY, USA), p. 146–161, Association for Computing Machinery, 2023.

[6] P. Hutchings, "Talking drums: Generating drum grooves with neural networks.," *CoRR*, vol. abs/1706.09558, 2017.

[7] D. Makris, M. Kaliakatsos-Papakostas, I. Karydis, and K. L. Kermanidis, "Conditional neural sequence learners for generating drums' rhythms," *Neural Comput. Appl.*, vol. 31, p. 1793–1804, jun 2019.

[8] H. F. Olson, *Music, Physics and Engineering*. Mineola, NY: Dover Publications, Sept. 1967.

[9] E. Britannica, "Harmony," 2024. Consulted on November 16, 2024.

[10] E. Britannica, "Melody," 2024. Consulted on November 16, 2024.

[11] SemiColonWeb, "Biblioteca digital." `https://encyclopediabritanica.uam.elogim.com/levels/collegiate/article/rhythm/110122`. Accessed: 2024-11-16.

[12] G. Loy, *Musimathics: V. 1.* London, England: MIT Press, July 2006.

[13] V. Lazzarini, S. Yi, J. ffitch, J. Heintz, Ø. Brandtsegg, and I. McCurdy, *MIDI Input and Output*, pp. 171–179. Cham: Springer International Publishing, 2016.

[14] The MIDI Association, "About MIDI - Part 3: MIDI messages," n.d. Accessed: 2024-11-14.

[15] M. Müller, *Music Representations*, pp. 1–38. Cham: Springer International Publishing, 2021.

[16] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," MM '20, (New York, NY, USA), p. 1180–1188, Association for Computing Machinery, 2020.

[17] F. Tirro, *Historia del Jazz (30 Aniversario)*. Redbook, Mar. 2024.

[18] J. A. Strain, *A Dictionary for the Modern Percussionist and Drummer*. Dictionaries for the Modern Musician, Lanham, Maryland: Rowman & Littlefield, 2017.

[19] N. Weinberg, "Guidelines for drumset notation," *Percussive Notes*, vol. 32, pp. 15–26, June 1994.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[21] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," 2017.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI*, 2019.

[24] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," in *NeurIPS EMC^{2}Workshop*, 2019.

[25] L. Moysis, L. A. Iliadis, S. P. Sotiroudis, A. D. Boursianis, M. S. Papadopoulou, K.-I. D. Kokkinidis, C. Volos, P. Sarigiannidis, S. Nikolaidis, and S. K. Goudos, "Music deep learning: Deep learning methods for music signal processing—a review of the state-of-the-art," *IEEE Access*, vol. 11, pp. 17031–17052, 2023.

[26] J. M. Simões, P. Machado, and A. C. Rodrigues, "Deep learning for expressive music generation," in *Proceedings of the 9th International Conference on Digital and Interactive Arts*, ARTECH '19, (New York, NY, USA), Association for Computing Machinery, 2020.

[27] Z. Li, Q. Huang, X. Yang, Q. Chen, and L. Zhang, "Automatic composition system based on transformer-xl," *Applied Sciences*, vol. 14, no. 13, 2024.

[28] Z. Yin, F. Reuben, S. Stepney, and T. Collins, "Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation," *Machine Learning*, vol. 112, pp. 1785–1822, May 2023.

[29] P. Pechimuthu, R. Ayup, P. Kamaraj, and S. A, "Automated creation of music using deep learning," in *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 1779–1782, 2024.

[30] M. A. Pangestu and S. Suyanto, "Generating music with emotion using transformer," in *2021 International Conference on Computer Science and Engineering (IC2SE)*, vol. 1, pp. 1–6, 2021.

[31] J. Min, Z. Gao, and L. Wang, "Application and research of music generation system based on cvae and transformer-xl in video background music," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 2, pp. 1409–1418, 2025.

[32] W. Wang, X. Li, C. Jin, D. Lu, Q. Zhou, and Y. Tie, "Cps: Full-song and style-conditioned music generation with linear transformer," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2022.

[33] N. Zhang, "Learning adversarial transformer for symbolic music generation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 4, pp. 1754–1763, 2023.

[34] C. Bao and Q. Sun, "Generating music with emotions," *IEEE Transactions on Multimedia*, vol. 25, pp. 3602–3614, 2023.

[35] Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Müller, and Y.-H. Yang, "Theme transformer: Symbolic music generation with theme-conditioned transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 3495–3508, 2023.

[36] J. Saputra, A. Prasetiadi, and I. Kresna, "Acapella-based music generation with sequential models utilizing discrete cosine transform," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 3, 2024.

[37] Y. Zhang, X. Lv, Q. Li, X. Wu, Y. Su, and H. Yang, "An automatic music generation method based on rscln_transformer network," *Multimedia Systems*, vol. 30, p. 4, Jan 2024.

[38] W. Huang, Y. Yu, H. Xu, Z. Su, and Y. Wu, "Hyperbolic music transformer for structured music generation," *IEEE Access*, vol. PP, pp. 1–1, 01 2023.

[39] Raffel, Colin and Ellis, Daniel P. W., "Intuitive Analysis, Creation and Manipulation of MIDI Data with pretty$_m idi,''$ $in 15th International Conference on Music Information Retrieval Late Breaking and Dem$

[40] N. Fradet, J.-P. Briot, F. Chhel, A. El Fallah Seghrouchni, and N. Gutowski, "MidiTok: A python package for MIDI file tokenization," in *Extended Abstracts for the Late-Breaking Demo Session of the 22nd International Society for Music Information Retrieval Conference*, 2021.

[41] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.

[42] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush, and T. Wolf, "Datasets: A community library for natural language processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online and Punta Cana, Dominican Republic), pp. 175–184, Association for Computational Linguistics, Nov. 2021.

[43] M. Koo and S.-W. Yang, "Likert-type scale," *Encyclopedia*, vol. 5, no. 1, 2025.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00118
Matrícula: 2232800209

GENERACIÓN DE SOLOS DE BATERÍA DE JAZZ CON HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL.

RODRIGO ALVARADO CASTAÑEDA
ALUMNO

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

En la Ciudad de México, se presentaron a las 15:00 horas del día 3 del mes de octubre del año 2025 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. ALFONSO PRIETO GUERRERO
DR. ELI GABRIEL AVIÑA BRAVO
DR. PEDRO LARA VELAZQUEZ

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN CIENCIAS (CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN)

DE: RODRIGO ALVARADO CASTAÑEDA

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

DIRECTOR DE LA DIVISIÓN DE CBI

Roman Linares Romero
DR. ROMAN LINARES ROMERO

PRESIDENTE

DR. ALFONSO PRIETO GUERRERO

VOCAL

DR. ELI GABRIEL AVIÑA BRAVO

SECRETARIO

DR. PEDRO LARA VELAZQUEZ