Ajuste de modelos de difusión para la generación de audio

Santiago Fiorino[1] and Pablo Riera[1,2]

[1] Departamento de Computación, Facultad de Ciencias Exactas y Naturalez, UBA, sfiorino@dc.uba.ar,
[2] Instituto de Ciencias de la Computación (UBA-CONICET)
Project Web: https://huggingface.co/santifiorino/SAO-Instrumental-Finetune

Abstract. La música ha evolucionado junto con la tecnología, desde instrumentos primitivos hasta modernas herramientas de síntesis digital. Actualmente, la inteligencia artificial desempeña un rol importante en la generación musical, empleando transformers y técnicas de difusión para crear canciones completas a partir de indicaciones en lenguaje natural. Sin embargo, los modelos privados, como los de Udio y Suno AI, aunque prometedores, limitan la investigación por su naturaleza cerrada. En 2024, Stability AI lanzó Stable Audio Open (SAO), un modelo de síntesis de audio basado en difusión y código abierto, democratizando el campo. Pese a su calidad en efectos de sonido, SAO enfrenta limitaciones en generación musical debido a datos de entrenamiento escasos y con licencias abiertas.

Nuestra investigación mejora las capacidades musicales de SAO mediante reentrenamiento con un conjunto de datos especializado. Se creó un pipeline que sintetiza audio a partir de archivos MIDI, enriquece metadatos con APIs (Spotify, LastFM) y genera indicaciones en lenguaje natural usando LLMs, obteniendo un conjunto de datos de 9 horas (538 minutos) compuesto por 1023 audios. Este conjunto incluye subconjuntos monofónico, polifónico y audios instrumentales de YouTube en partes iguales, con variedad de géneros, tempos, y tonalidades para diversificar la sonoridad. El modelo reentrenado ("Instrumental Finetune") supera al SAO original, logrando mejoras en calidad de sonido, precisión instrumental y adherencia a géneros y tempos, alcanzando un 95,3% de precisión frente al 77,6% original. Aunque los desafíos en tonalidad persisten, métricas como KL-Passt y CLAP Score muestran que nuestro modelo iguala o supera el rendimiento de SAO y MusicGen, manteniendo generalización y optimización específica del dominio. Ejemplos auditivos que ilustran estas mejoras y confirman la ausencia de memorización están disponibles en el Project Web.

Keywords: música, síntesis, difusión, *transformers*

Fine-Tuning Diffusion Models for Audio Generation

Abstract. Music has evolved alongside technological advancements, from primitive percussion to modern digital synthesis tools. Today, artificial intelligence plays important role in music generation, utilizing state-of-the-art architectures like transformers and diffusion models to generate complete songs from natural language prompts. Proprietary models by Udio and Suno AI demonstrate great potential but limit scientific research due to their closed nature. In June 2024, Stability AI released Stable Audio Open (SAO), an open-source diffusion-based audio synthesis model, democratizing research in this field. While SAO excels in sound effect generation, its musical capabilities are limited by scarce open-license training data.

Our research enhances SAO's musical generation capabilities through fine-tuning on a specialized dataset, addressing its inability to generate certain instruments, difficulties with specified musical elements, and inconsistencies in tempo and tonality. We developed a custom dataset-creation pipeline by synthesizing audio from MIDI files, enriching metadata using APIs like Spotify and LastFM, and generating natural language prompts via large language models. This pipeline produced a 9-hour (538 minutes) music dataset comprising 1023 audios, which includes monophonic, polyphonic, and instrumental YouTube audio subsets in equal parts, spanning various genres, tempos, and tonalities. Results show significant improvements in the fine-tuned model ("Instrumental Finetune") over the original SAO, particularly in sound quality, instrument reproduction accuracy, genre adherence, and tempo adherence (95.3% accuracy vs. 77.6%). Although tone and scale accuracy remain challenging, embedding-based metrics (KL-Passt, CLAP Score) indicate our model matches or surpasses both SAO and the commercial MusicGen, maintaining generalization despite domain-specific optimization. Auditory examples illustrating these improvements and confirming the absence of memorization are available on the Project Web.

Keywords: music, synthesis, diffusion, transformers

1 Introducción

La generación de música mediante inteligencia artificial ha emergido como un campo de investigación y de aplicación con gran potencial artístico. Los avances recientes en arquitecturas de redes neuronales, particularmente los transformers y los modelos de difusión, han permitido la creación de sistemas capaces de sintetizar composiciones musicales complejas a partir de simples indicaciones en lenguaje natural. Si bien modelos privados como los desarrollados por Udio y Suno AI han demostrado capacidades impresionantes, su naturaleza propietaria dificulta la investigación y el avance en el ámbito académico. En este contexto, en junio de 2024 lanza Stable Audio Open (SAO), un modelo de síntesis de audio basado en difusión y de código abierto, representa un hito crucial al democratizar la investigación en este campo [Evans, Parker, et al., 2024]. SAO destaca en la generación de efectos de sonido, pero sus capacidades para la síntesis musical se ven limitadas por la disponibilidad de datos de entrenamiento musicales con licencias abiertas. Esta limitación motiva nuestra investigación, cuyo objetivo principal es mejorar las capacidades de generación musical de SAO a través del re-entrenamiento con un conjunto de datos especializado.

2 Breve historia de la síntesis de audio

La síntesis de audio ha experimentado una importante evolución a lo largo de más de un siglo, desde los instrumentos electrónicos de principios del siglo XX hasta los sintetizadores analógicos, la revolución digital y la era del software con los sintetizadores virtuales y las estaciones de trabajo de audio digital (DAWs). Cada etapa ha marcado avances significativos en la portabilidad, el control, la expresividad y la calidad sonora de los instrumentos musicales electrónicos. Paralelamente, emergieron métodos para la composición automática de música. Inicialmente, estos modelos seguían reglas definidas manualmente o usaban cadenas de Markov [Westergaard and Hiller, 1959; Zaripov, 1960]. Algunos intentaron no solo componer música, sino replicar estilos específicos a partir de conjuntos de datos [Coenen, 1997; Pachet, 2010]. Más recientemente, el auge de la inteligencia artificial redefinió el campo, comenzando con redes Long Short-Term Memory (LSTM), que superaron la limitación de coherencia a largo plazo de las redes recurrentes (RNNs) [Eck and Schmidhuber, 2002]. Posteriormente, las redes convolucionales introdujeron modelos capaces de sintetizar directamente el audio final [van den Oord et al., 2016]. Entre los avances más recientes destacan los transformers [Sohl-Dickstein et al., 2015], que tras sus éxitos en procesamiento de texto e imágenes, se aplicaron a la síntesis musical [Huang et al., 2018]. Además, el método de difusión, destacado inicialmente en la generación de imágenes [Ho et al., 2020; Sohl-Dickstein et al., 2015], se adaptó rápidamente a la síntesis de audio [Evans, Carr, et al., 2024; Forsgren and Martiros, 2022]. Los modelos de difusión con arquitecturas de transformers son actualmente los sistemas de síntesis *end-to-end* más avanzados, junto con los modelos de transformers puros, autoregresivos, que ofrecen alta calidad aunque con tiempos de inferencia significativamente mayores [Copet et al., 2023].

3 Bases técnicas fundamentales

El funcionamiento de Stable Audio Open se fundamenta en la combinación de tres pilares técnicos: los autoencoders variacionales (VAEs), los modelos de difusión y las arquitecturas transformer. El VAE se encarga de aprender una representación latente comprimida del audio, permitiendo que el modelo opere en un espacio de menor dimensionalidad. El proceso de difusión consiste en aprender a revertir un proceso gradual de adición de ruido a los datos, lo que permite generar nuevas muestras partiendo de ruido aleatorio. Los transformers, basados en mecanismos de atención, posibilitan modelar el proceso de difusión, que originalmente se modelaba con redes U-Net [Peebles and Xie, 2023]. SAO utiliza un VAE que opera directamente sobre las ondas de audio y un transformer de difusión para refinar las representaciones latentes. El VAE reduce la dimensión temporal del audio crudo en un factor de 1024, logrando una compresión de 32x. Su transformer de difusión, basado en arquitecturas ViT, condiciona las generaciones mediante embeddings de texto (T5-base), el timing y el paso de difusión, usando capas de self-attention y cross-attention. La capacidad de condicionar el proceso de difusión mediante técnicas como el classifier-free guidance [Ho and Salimans, 2022], utilizando embeddings de texto, permite guiar la generación de audio hacia el contenido deseado, lo que resulta fundamental para la síntesis musical controlada por prompts en lenguaje natural.

4 Desarrollo de la investigación

4.1 Creación del conjunto de datos especializado

Una de las partes centrales de la investigación fue la creación de un conjunto de datos especializado para el re-entrenamiento de SAO con el objetivo de mejorar sus capacidades musicales. Dada la dificultad de obtener grandes cantidades de audio musical de alta calidad y con metadatos detallados bajo licencias abiertas, optamos por una metodología de síntesis de audio a partir de archivos MIDI. Este enfoque nos brindó control total sobre el contenido generado y la posibilidad de obtener metadatos precisos. El proceso de creación del conjunto de datos comprendió las siguientes etapas:

Selección y limpieza de archivos MIDI: Se utilizó el "Clean MIDI subset" del Lakh MIDI Dataset v0.1. Se implementaron procesos para eliminar archivos duplicados y aquellos que no pudieron ser abiertos, resultando en un conjunto depurado de archivos MIDI.

Síntesis de audio: Cada archivo MIDI fue separado en pistas individuales por instrumento y renderizado utilizando instrumentos virtuales (VSTs) a través de la biblioteca pedalboard de Spotify. Se aleatorizaron parámetros de los VSTs para generar variaciones sonoras. Las pistas renderizadas fueron normalizadas en volumen utilizando la recomendación ITU-R BS.1770 y mezcladas en un único archivo de audio. Finalmente, los audios se segmentaron en secciones más cortas basadas en la segmentación proporcionada por la API de Spotify.

Generación de prompts en lenguaje natural: Se enriquecieron los metadatos de cada audio utilizando las APIs de Spotify y LastFM. De Spotify se obtuvieron características musicales técnicas y perceptuales, así como la segmentación de las canciones. De LastFM se obtuvieron etiquetas descriptivas generadas por usuarios, que complementaron la descripción de los géneros y estilos musicales. Se eliminaron etiquetas relacionadas con vocales, ya que nuestros renders eran puramente instrumentales. Se realizaron predicciones independientes de tempo, tonalidad y escala utilizando herramientas como essentia. Finalmente, se utilizó un modelo de lenguaje grande (LLM) para transformar los metadatos estructurados en prompts descriptivos en lenguaje natural. Se implementaron técnicas de prompts dinámicos y few-shot learning para mejorar la diversidad y precisión de los prompts generados.

Aplicación web para curación: Se desarrolló una aplicación web para visualizar, editar y filtrar los pares ⟨audio, prompt⟩ generados. Esta herramienta permitió la inspección auditiva de los audios, la edición de los prompts y la corrección de la lista de instrumentos presentes en cada sección de audio, solucionando inconsistencias generadas por la segmentación.

Mediante este proceso se creó un conjunto de datos de 9 horas de música que abarcaba diversos géneros, tempos y tonalidades. El conjunto final se dividió en subconjuntos monofónico, polifónico y uno adicional con audios instrumentales sin derechos de autor de YouTube para diversificar la sonoridad.

5 Re-entrenamiento del modelo Stable Audio Open

Con el conjunto de datos curado, se procedió al re-entrenamiento del modelo Stable Audio Open. El entrenamiento se realizó en un entorno Google Colab con una tarjeta gráfica A100. Se utilizaron los scripts de entrenamiento proporcionados por Stability AI, ajustando parámetros como el batch size, que se incrementó progresivamente hasta 16. El entrenamiento se extendió por un total de 5 horas, completando 4000 pasos en 63 épocas. El modelo resultante de este proceso se denominó “Instrumental Finetune”.

6 Resultados y evaluación

Para evaluar el modelo se realizaron análisis auditivos y de métricas objetivas:

Evaluación auditiva: Se generaron audios a partir de prompts nuevos y del conjunto de entrenamiento, comparándolos con el modelo SAO original. Se observó una notable mejora en la calidad del sonido, mezcla, definición y aspectos melódicos. El modelo “Instrumental Finetune” mostró mejor capacidad para sintetizar instrumentos que el original no generaba correctamente, como trompetas y saxofones, además de una mayor adherencia a los géneros solicitados en los prompts. No se detectaron casos de memorización del conjunto de entrenamiento.

Ajuste al tempo: Se generaron audios con prompts que especificaban tempos en un rango de BPM y se estimó el tempo utilizando deeprhythm [Aarabi

and Peeters, 2019]. Se aplicaron las métricas Accuracy 1 y Accuracy 2 [Gouyon et al., 2006]. Accuracy 1 es el porcentaje de predicciones dentro de un 2% del tempo real del audio. Accuracy 2 incluye también predicciones dentro de un 2% del doble, triple, mitad o tercio del tempo. La tabla 1 muestra una mejora significativa en la adherencia al tempo con el modelo "Instrumental Finetune".

Ajuste a la tonalidad y escala: Se generaron audios para cada combinación de nota y escala, y se predijeron estos parámetros utilizando essentia. En contraste con el tempo, ambos modelos mostraron una baja adaptación a la tonalidad y escala requeridas. Sin embargo, la matriz de confusión del modelo re-entrenado exhibió una distribución de aciertos más equilibrada. En la tabla 1 se puede ver que el accuracy en la predicción conjunta de nota y escala mejoró ligeramente en el modelo re-entrenado, sugiriendo una mayor precisión en la escala cuando la nota generada era la esperada.

Métricas basadas en representaciones: Se utilizaron las métricas KL-passt y CLAP_{score}, calculadas con el repositorio `stable-audio-metrics`, para comparar las generaciones de "Instrumental Finetune" y SAO en un subconjunto del Song Describer Dataset. Los resultados indicaron que el modelo re-entrenado mantuvo o superó el rendimiento de SAO en estas métricas, sugiriendo que el re-entrenamiento en un conjunto de datos especializado no deterioró las capacidades generales del modelo.

	Acc. 1	Acc. 2	Acc. nota y escala	KL_{passt} ↓	CLAP_{score} ↑
Stable Audio Open (SAO)	0.747	0.776	0.21	0.54	0.39
SAO Instrumental Finetune	0.876	0.953	0.27	0.52	0.38

Table 1. Métricas de Accuracy 1, Accuracy 2, Accuracy de nota y escala, KL_{passt} y CLAP_{score} para Stable Audio Open (SAO) e Instrumental Finetune, sobre 200 generaciones de audios a partir de prompts que nunca antes vieron.

7 Conclusiones

Esta investigación demostró la viabilidad de mejorar significativamente las capacidades de generación musical del modelo de difusión de código abierto Stable Audio Open mediante el re-entrenamiento con un conjunto de datos especializado. El pipeline desarrollado para la creación automática de música y prompts, la curación manual del conjunto de datos y el proceso de re-entrenamiento resultaron en un modelo ("Instrumental Finetune") con calidad de sonido, precisión instrumental y adherencia a géneros y tempo superiores al modelo original. Si bien la adaptación a la tonalidad y la escala sigue siendo un desafío, los resultados generales son prometedores y abren nuevas vías para la investigación en la generación de audio musical con modelos de difusión de código abierto. La capacidad de mantener las capacidades de generalización del modelo original al mismo tiempo que se especializa en un dominio específico es un hallazgo relevante.

References

Aarabi, H. F., & Peeters, G. (2019). Deep-rhythm for global tempo estimation in music. *ISMIR*, 636–643.

Coenen, A. (1997). David cope, experiments in musical intelligence. a-r editions, madison, wisconsin, usa. vol. 12 1996. *Organised Sound*, *2*(1), 57–60. https://doi.org/10.1017/S1355771897210101

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). Simple and controllable music generation.

Eck, D., & Schmidhuber, J. (2002). A first look at music composition using lstm recurrent neural networks.

Evans, Z., Carr, C., Taylor, J., Hawley, S. H., & Pons, J. (2024). Fast timing-conditioned latent audio diffusion. https://arxiv.org/abs/2402.04825

Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J., & Pons, J. (2024). Stable audio open. https://arxiv.org/abs/2407.14358

Forsgren, S., & Martiros, H. (2022). Riffusion - Stable diffusion for real-time music generation. https://riffusion.com/about

Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., & Cano, P. (2006). An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1832–1844. https://doi.org/10.1109/TSA.2005.858509

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. https://arxiv.org/abs/2006.11239

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. https://arxiv.org/abs/2207.12598

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music transformer. https://arxiv.org/abs/1809.04281

Pachet, F. (2010). The continuator: Musical interaction with style. *Journal of New Music Research*, *32*, 333–341. https://doi.org/10.1076/jnmr.32.3.333.16861

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. https://arxiv.org/abs/2212.09748

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. https://arxiv.org/abs/1503.03585

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. https://arxiv.org/abs/1609.03499

Westergaard, P., & Hiller, L. A. (1959). *Journal of Music Theory*, *3*(2), 302–306. Retrieved November 15, 2024, from http://www.jstor.org/stable/842857

Zaripov, R. K. (1960). An algorithmic description of a process of musical composition. *Dokl. Akad. Nauk SSSR*, *132*, 1283–1286. http://mi.mathnet.ru/dan23732