FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN CIENCIA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2024/2025

Trabajo de Fin de Máster

TITULO: Técnica de Generación de Audios Sintéticos Multilenguaje y Multiacento con Control de las Características de la Voz

***Alumno*: Sara Vanesa Orozco Narváez**

***Tutor*: Javier Portela García-Miguel**
Ana Lucila Sandoval Orozco

Julio de 2025

UNIVERSIDAD COMPLUTENSE
MADRID

Declaración Responsable sobre Autoría y Uso Ético de Herramientas de Inteligencia Artificial (IA)

Yo, Sara Vanesa Orozco Narvaez

Con DNI/: 60817502M, declaro de manera responsable que el/la presente:

Trabajo de Fin de Máster (TFM)

Titulado/a

Técnica de Generación de Audios Sintéticos Multilenguaje y Multiacento con Control de las Características de la Voz

es el resultado de mi trabajo intelectual personal y creativo, y ha sido elaborado de acuerdo con los principios éticos y las normas de integridad vigentes en la comunidad académica y, más específicamente, en la Universidad Complutense de Madrid.

Soy, pues, autor del material aquí incluido y, cuando no ha sido así y he tomado el material de otra fuente, lo he citado o bien he declarado su procedencia de forma clara -incluidas, en su caso, herramientas de inteligencia artificial-. Las ideas y aportaciones principales incluidas en este trabajo, y que acreditan la adquisición de competencias, son mías y no proceden de otras fuentes o han sido reescritas usando material de otras fuentes.

Asimismo, aseguro que los datos y recursos utilizados son legítimos, verificables y han sido obtenidos de fuentes confiables y autorizadas. Además, he tomado medidas para garantizar la confidencialidad y privacidad de los datos utilizados, evitando cualquier tipo de sesgo o discriminación injusta en el tratamiento de la información.

En Madrid a Junio 19, 2025

OROZCO NARVAEZ SARA VANESA - DNI Y9266939T

Digitally signed by OROZCO NARVAEZ SARA VANESA - DNI Y9266939T
Date: 2025.06.19 14:25:56 +02'00'

FIRMA

Agradecimientos

Quisiera expresar mi más sincero agradecimiento a mis tutores de este Trabajo de Fin de Máster, por su constante apoyo, orientación y confianza durante todo el proceso. Su experiencia, compromiso y disponibilidad han sido fundamentales para el desarrollo de este proyecto.

También quiero agradecer a mis compañeros del máster, con quienes compartí incontables horas de estudio, reuniones y desafíos. Desde las clases entre semana hasta encuentros de trabajo que se extendían incluso a los fines de semana, siempre encontramos la manera de ayudarnos y apoyarnos mutuamente. Su compañerismo y compromiso hicieron que esta etapa fuera no solo formativa, sino también profundamente enriquecedora a nivel personal. Me alegra saber que, más allá del ámbito académico, se convirtieron en verdaderos amigos.

Asimismo, agradezco a todos los profesores del máster por su dedicación, esfuerzo y pasión por la enseñanza. Su compromiso ha sido clave para ampliar mis conocimientos y despertar mi interés por áreas que hoy forman parte de mi trayectoria profesional.

Índice General

Índice de Figuras

Índice de Tablas

Lista de Acrónimos

AR	*Modelos Autorregresivos*
CC	*Coeficientes Cepstrales*
CNN	*Redes Neuronales Convolucionales*
CSS	*Síntesis de Voz Concatenativa*
DCT	*Transformada Discreta del Coseno*
DL	*Aprendizaje Profundo*
FAD	*Frechet Audio Distance*
FD	*Frechet Distance*
GAN	*Redes Generativas Adversarias*
IA	*Inteligencia Artificial*
KL	*Kullback-Leibler Divergence*
LDM	*Modelos de Difusión Latente*
LLM	*Modelos de Lenguaje Grandes*
LSTM	*Modelos de Memoria a Largo Plazo*

MFCCs	*Coeficientes Cepstrales de Frecuencia Mel*
ML	*Aprendizaje Automático*
MOS	*Puntuación de opinión media*
PSS	*Síntesis de Voz Paramétrica*
RBM	*Máquinas de Boltzmann Restringidas*
RNN	*Redes Neuronales Recurrentes*
SPSS	*Síntesis de Voz Paramétrica Estadística*
STFT	*Transformada de Fourier de Corto Tiempo*
T2I	*Texto a Imagen*
TTA	*Texto a Audio*
TTS	*Texto a Voz*
VAE	*Codificadores Automáticos Variacionales*
VQ-VAE	*Autocodificador Variacional Cuantificado Vectorial*
WER	*Word Error Rate*

Abstract

In recent years, synthetic audio generation has advanced significantly due to continuous developments and innovations in deep learning models. As a result, the demand for models capable of generating realistic and high-quality synthetic audio continues to grow, and generating synthetic audio with specific characteristics such as age, gender, language, and accent remains a challenging task. To address these challenges, this work proposes an approach for synthetic audio generation that combines the VoiceLDM and XTTS models to produce enhanced synthetic audio with multilingual and multi-accent capabilities. The proposed approach follows a two-phase process for generating synthetic audio. First, the VoiceLDM model is used to generate the base audio, assigning specific voice characteristics such as age, gender, language, and accent. The base audio is generated in English, Spanish, and Portuguese. To address this limitation and expand audio generation to other languages, the second phase uses the XTTS model, which takes the VoiceLDM output and generates new versions of the audio, either preserving the original language or producing new audios in other languages, while ensuring voice consistency and speaker characteristics. Experimental results demonstrate the effectiveness of the proposed approach, showing improvements in the quality and accuracy of the generated audio. In particular, a Word Error Rate (WER) of 18,5 % was achieved—the lowest among all evaluated models—indicating high linguistic fidelity. Additionally, it achieved the highest CLAP Score (0,229), very close to that of real audio (0,273), reflecting better semantic alignment between the text and the generated audio. Moreover, the obtained values of Frechet Audio Distance (FAD = 2,082) and Kullback-Leibler Divergence (KL = 0,003) indicate that the generated audios are very similar to the real reference audios. Taken together, these results show that the VoiceLDM + XTTS approach achieves an outstanding balance between acoustic quality, semantic coherence, and linguistic fidelity, outperforming state-of-the-art models in personalized synthetic audio generation tasks with multilingual and multi-accent capabilities.

Keywords: Deep learning, Text-to-speech, Text-to-audio, Voice style control, Multilingual voice synthesis, Synthetic audio generation, Voice characteristics.

Resumen

En los últimos años, la generación de audio sintético ha avanzado significativamente debido a los continuos desarrollos y las innovaciones en modelos de aprendizaje profundo. Como resultado, la necesidad de modelos que generen audio sintético realista y de alta calidad sigue en aumento y generar audios sintéticos con características específicas como edad, género, idioma y acento sigue siendo una tarea desafiante. Para abordar estos retos, este trabajo propone un enfoque para la generación de audios sintéticos que combina los modelos VoiceLDM y XTTS para obtener un audio sintético mejorado con capacidades multilingües y multi-acento. El enfoque propuesto sigue un proceso de dos fases para la generación del audio sintético. Primero se utiliza el modelo VoiceLDM para generar el audio base, asignando características específicas a la voz como edad, género, idioma y acento. Los idiomas en los que se genera el audio base son inglés, español y portugués. Para abordar esta limitación y ampliar la generación de audio a otros idiomas, en la segunda fase se utiliza el modelo XTTS, que toma la salida de VoiceLDM y genera nuevas versiones del audio, manteniendo el idioma original o generando nuevos audios en otros idiomas, asegurando la coherencia de la voz y las características del hablante. Los resultados experimentales demuestran la efectividad del enfoque propuesto, reflejando mejoras en la calidad y precisión del audio generado. En particular, se alcanzó un Word Error Rate (WER) de 18,5 %, el más bajo entre todos los modelos evaluados, lo que indica una alta fidelidad lingüística. Asimismo, obtuvo el CLAP Score más alto (0,229), muy cercano al del audio real (0,273), reflejando una mejor alineación semántica entre el texto y el audio generado. Además, los valores obtenidos de Frechet Audio Distance (FAD = 2,082) y Kullback-Leibler Divergence (KL = 0,003) evidencia que los audios generados son muy similares a los audios reales de referencia. En conjunto, estos resultados muestran que el enfoque VoiceLDM + XTTS logra un equilibrio sobresaliente entre calidad acústica, coherencia semántica y fidelidad lingüística, superando a modelos del estado del arte en tareas de generación de audio sintético personalizado con capacidad multilingüe y multiacento.

Palabras clave: Aprendizaje profundo, Texto a voz, Texto a Audio, Control de estilos de voz, Síntesis de voz multilingüe, Generación de audio sintético, Características de voz.

Capítulo 1

Introducción

En este capítulo se presenta una introducción general al trabajo realizado, abordando la motivación detrás de la investigación sobre la generación de audio sintético con capacidades multilingüe y multiacento, como se detalla en la Sección 1.1. La Sección 1.2 contextualiza el trabajo dentro del marco del proyecto europeo *ALUNA*, que se centra en el desarrollo de estrategias tecnológicas, metodológicas y sociales para la prevención, investigación y asistencia a las víctimas de los crímenes de abuso y explotación sexual infantil. En la Sección 1.3, se definen el objetivo general y los objetivos específicos, centrados en la integración de los modelos VoiceLDM y XTTS con un discriminador basado en Wav2Vec 2.0 para superar las limitaciones actuales en el control de características vocales. Además, en la Sección 1.4, se describe la estructura del documento, proporcionando una visión general de los capítulos que desarrollan el marco teórico, el estado del arte, la metodología, los experimentos y las conclusiones.

1.1. Motivación

En los últimos años, el interés por la generación de audio sintético ha ido en aumento, transformando la forma en la que interactuamos con la tecnología, desempeñando un papel muy importante en diversas aplicaciones que requieren voces realistas y personalizadas. Tecnologías como la síntesis de Texto a Voz (Texto To Speech TTS) y Texto a Audio (Text To Audio TTA) han permitido la creación de sistemas capaces de convertir texto en voz con una calidad cada vez más cercana a la del habla humana. Estos avances han facilitado el desarrollo de asistentes virtuales, lo cuales permiten la interacción natural con dispositivos como Siri, Google Assistant y Alexa [AR24]. En accesibilidad, ayuda a personas con discapacidad visual mediante lectores de pantalla. En navegación, convierte instrucciones de texto en voz en dispositivos GPS. En educación, facilita el acceso a contenido mediante audiolibros y plataformas de aprendizaje. En centros de llamadas, automatiza respuestas para mejorar la atención al cliente. En aplicaciones multilingües, apoya la traducción y el aprendizaje de idiomas. En entretenimiento, permite la creación de audiolibros y podcasts.

Los algoritmos de aprendizaje profundo están siendo implementados en la generación de audio sintético, impulsando progresos importantes en la síntesis de texto a voz (TTS), facilitando la creación de habla con más naturalidad, fluidez y expresividad [WCM+21]. Los métodos convencionales de TTS, como los modelos concatenativos y paramétricos, están siendo reemplazados por arquitecturas basadas en redes neuronales profundas, lo que ha permitido mejorar la calidad del habla sintetizada al capturar características prosódicas y fonéticas más complejas.

Los modelos actuales de TTS emplean estructuras basadas en aprendizaje profundo, tales como *Redes Generativas Adversarias* (GAN), Transformers y modelos de difusión latente, con el fin de optimizar la calidad del audio generado. Modelos como Tacotron [WSRS+17], FastSpeech [RRT+19] y Deep Voice 3 [PPG+18] son ejemplos de estos avances, que mejoran la alineación entre texto y audio, disminuyendo los errores en la síntesis de voz [RRT+19, Łań21, CDG+24]. Adicionalmente, los vocoders neuronales, como HiFi-GAN y WaveGlow, han facilitado la conversión de representaciones intermedias en formas de onda de alta calidad, mejorando el realismo del audio generado [KPK+23, Bet23].

A pesar de estos progresos, la síntesis de voz sigue enfrentando importantes desafíos. La generación de audios sintéticos con características específicas, como edad, género, idioma y acento, continúa siendo un reto debido a la dificultad de capturar y replicar con precisión las variaciones prosódicas, fonéticas y el estilo del habla humana. Además, muchos modelos actuales dependen de grandes volúmenes de datos etiquetados para lograr una calidad aceptable, lo que limita su aplicabilidad en idiomas con pocos recursos o en escenarios donde se requiere una personalización avanzada de la voz [AR24].

Muchos modelos en la actualidad permiten generar audios con control del estilo de la voz, pero limitan las características del hablante a aspectos como el género, tono, timbre y emoción. Pero no todos son multilingües ni permiten generar audios por rangos de edades. Modelos como XTTS [CDG+24] están diseñados para ser multilingües, permitiendo generar voz en varios idiomas sin necesidad de reentrenamiento específico, sin embargo, no todos los modelos que controlan el estilo de la voz son multilingües. Por otro lado, modelos como VoiceLDM [LYNC24] pueden generar voces que suenan naturales y que se pueden controlar sus características definiendo el contexto ambiental en el que se va a generar el audio. Sin embargo, no es posible especificar rangos de edad exactos para la voz generada, limitándolo a descripciones generales como “una mujer joven” o “un anciano”. Por lo tanto, la capacidad de generar audios en rangos de edad específicos y en múltiples idiomas aún no está ampliamente implementado en la mayoría de los modelos actuales.

Para abordar estas limitaciones, en este trabajo se propone un enfoque innovador para la generación de audio sintético en dos fases, integrando los modelos VoiceLDM y XTTS. En la primera fase, se utiliza VoiceLDM para generar un audio base en inglés, español o portugués, permitiendo definir características específicas de la voz como el rango de edad, el género y el acento. En la segunda fase, se emplea XTTS, que recibe la salida de VoiceLDM y genera nuevas versiones del audio en diferentes idiomas, preservando la

identidad vocal del hablante.

Además, para garantizar que el audio generado corresponda con las características deseadas del hablante proporcionadas en la entrada del modelo, como el rango de edad y el género, se integra un mecanismo de validación posterior a la generación del audio. Este mecanismo emplea un modelo discriminador entrenado específicamente para estimar edad y género a partir del audio. Si el resultado no se ajusta a los parámetros especificados, el proceso de generación se repite hasta obtener un audio que cumpla con los criterios definidos, asegurando así coherencia y fidelidad en las características vocales del resultado final.

1.2. Contexto de la Investigación

El presente Trabajo de Fin de Máster se enmarca dentro de un proyecto de investigación titulado *Child Protection Centred Strategies to Fight Against Sexual Abuse and Exploitation – ALUNA*. ALUNA ha recibido financiación del Fondo de Seguridad Interna de la Unión Europea en virtud del acuerdo de subvención no 101084929. y en el que participa como coordinador del proyecto el Grupo GASS de la Universidad Complutense de Madrid (Grupo de Análisis, Seguridad y Sistemas, https://gass.ucm.es, grupo 910623 del catálogo de grupos de investigación reconocidos por la UCM).

Además de la Universidad Complutense de Madrid – UCM participan en ALUNA 16 socios de 5 países de la UE (España, Bélgica, Grecia, Irlanda y Portugal), 2 Países Asociados (Suiza e Israel) y 3 Terceros países (Brasil, Colombia y Reino Unido). University of Kent (Reino Unido), The Free University of Brussels (Bélgica), Center for Security Studies – KEMEA (Grecia), International Center for Missing and Exploited Children – ICMEC (Suiza), IDENER Research y Development Agrupación de Interés Económico (España), Athena Research Center (Grecia), Trilateral Research and Consulting (Reino Unido), Portuguese Association for Victim Support – APAV (Portugal), Fundación Renacer (Colombia), Brazilian Association for the Defense of Children of Children and Youth – ASBRAD (Brazil), Hellenic Police (Grecia), Dirección General de la Policía – DGP (España), Federal Police (Brazil), Federal Highway Police (Brazil), Ministry of Public Security - Israel National Police – MOPS-INP (Israel)

1.3. Objetivos

Los objetivos de esta investigación se dividen en un objetivo general y tres objetivos específicos, orientados a desarrollar y validar un marco para la generación de audio sintético multilenguaje y multiacento a partir de descripciones textuales personalizadas y contenido lingüístico.

1.3.1. Objetivo General

Proponer un enfoque para la generación de audio sintético multilingüe y multiacento que integre los modelos VoiceLDM y XTTS, junto con un discriminador basado en Wav2Vec 2.0, con el fin de producir audios de alta calidad capaces de incorporar características específicas de la voz como edad, género, idioma y acento, superando las limitaciones de modelos anteriores que no permiten controlar simultáneamente estos atributos.

1.3.2. Objetivos Específicos

Para alcanzar el objetivo general, se plantean los siguientes objetivos específicos:

1. Ajustar y evaluar el modelo VoiceLDM mediante fine-tuning con datasets de audio limpio (Common Voice 18.0, LibriTTS, CML TTS), optimizando su capacidad para generar audio base que refleje descripciones textuales y contenido lingüístico con alta naturalidad y fidelidad.
2. Implementar y validar un discriminador basado en Wav2Vec 2.0 para verificar que el audio base generado por VoiceLDM cumpla con las especificaciones de edad y género definidas en la entrada, iterando el proceso de generación hasta lograr coherencia.
3. Integrar el modelo XTTS para adaptar el audio base a múltiples idiomas, preservando las características vocales del hablante.

1.4. Estructura del trabajo

Este trabajo se estructura en seis capítulos, organizados de manera lógica para abordar la propuesta de generación de audio sintético desde sus fundamentos teóricos hasta su implementación y evaluación experimental:

- En el **Capítulo 1**, se presenta la introducción general del trabajo, incluyendo la motivación del estudio, el objetivo general, los objetivos específicos y esta descripción de la estructura del documento.

- El **Capítulo 2** expone el **trasfondo sobre la investigación**, donde se desarrollan los conceptos teóricos fundamentales que sustentan el trabajo. Se abordan temas como la inteligencia artificial, los modelos generativos, el procesamiento de señales de audio, las representaciones acústicas, y las tecnologías de TTS y TTA.
- En el **Capítulo 3** se revisa el **estado del arte**, analizando los principales modelos en generación de audio sintético. Se exploran soluciones basadas en modelos de difusión, redes neuronales profundas, y técnicas para el control de voz multilingüe y multiacento.
- El **Capítulo 4** describe la **propuesta y metodología** del enfoque desarrollado. Se detalla el proceso de *fine-tuning* del modelo VoiceLDM con audio limpio, el flujo completo de generación de audio sintético y la integración con el modelo XTTS para hacer que sea multilenguaje.
- En el **Capítulo 5** se presentan los **experimentos y resultados**. Se describe la configuración del entorno, las métricas empleadas para la evaluación, los resultados obtenidos tras el ajuste del modelo VoiceLDM y la evaluación comparativa del enfoque propuesto frente a modelos del estado del arte.
- Finalmente, el **Capítulo 6** recoge las **conclusiones** del trabajo, destacando los aportes principales, y propone posibles líneas de trabajo futuro para mejorar y ampliar el sistema desarrollado.

1.5. Síntesis del capítulo

El objetivo de este capítulo es presentar una introducción general a la investigación sobre la generación de audio sintético multilingüe y multiacento, destacando su relevancia y los desafíos actuales en el campo de TTS y TTA. Se ha abordado la motivación del estudio, enfatizando la importancia de estas tecnologías en aplicaciones como accesibilidad, educación y entretenimiento, así como las limitaciones en el control de características vocales específicas, como edad y acento. Además, se han definido el objetivo general y los objetivos específicos, centrados en la integración de los modelos VoiceLDM y XTTS con un discriminador basado en Wav2Vec 2.0 para superar estas limitaciones. Finalmente, se ha descrito la estructura del documento, proporcionando un esquema claro de los capítulos que se desarrollan en este trabajo.

Capítulo 2

Marco Teórico

En este capítulo se desarrolla el marco teórico que sustenta la propuesta de generación de audio sintético, abordando los conceptos fundamentales desde una perspectiva general hasta los específicos del trabajo. En la Sección 2.1, se introduce la *Inteligencia Artificial* (IA), describiendo su evolución y aplicaciones. En la Sección 2.2, se exploran los modelos generativos, incluyendo redes generativas adversarias, codificadores automáticos variacionales, modelos de difusión latente y modelos autorregresivos. Posteriormente, en la Sección 2.3, se analiza el procesamiento de señales de audio, detallando las representaciones más comunes (como espectrogramas de Mel y embeddings acústicos) y técnicas de preprocesamiento de audio y texto. Finalmente, en las Secciones 2.4 y 2.5, se describen los sistemas de TTS y TTA, destacando sus arquitecturas, avances y desafíos, sentando las bases teóricas para la propuesta de generación de audio multilingüe y multiacento.

2.1. Inteligencia Artificial

La IA es una disciplina de la informática que busca desarrollar sistemas con la capacidad de simular funciones cognitivas humanas como el aprendizaje, el razonamiento, la percepción y la toma de decisiones, con el objetivo de adaptarse a diversos contextos y resolver tareas complejas de manera autónoma o asistida [Sha24]. Estos sistemas se basan en algoritmos y modelos computacionales que aprenden de grandes conjuntos de datos, permitiendo a las máquinas interpretar entornos, generalizar conocimientos y ejecutar acciones sin intervención humana directa en muchos casos.

La IA ha ido avanzando a lo largo de los años y han ido evolucionando los diferentes enfoques metodológicos a lo largo de su desarrollo. En sus inicios, se centraba en modelos simbólicos y basados en reglas, que empleaban lógica formal para representar el conocimiento y realizar inferencias [HQR+17].

Posteriormente, se desarrollaron enfoques basados en métodos estadísticos y de aprendizaje automático, donde los sistemas aprenden patrones a partir de datos en lugar de depender exclusivamente de reglas predefinidas. Estos métodos permitieron a las

máquinas generalizar comportamientos a partir de ejemplos, incrementando la flexibilidad y capacidad de adaptación ante entornos dinámicos [Sch22].

Más recientemente, el avance de las técnicas de aprendizaje profundo (deep learning) ha permitido grandes progresos en áreas como la visión por computador, el procesamiento del lenguaje natural y la generación de contenido sintético. Estas técnicas utilizan redes neuronales profundas capaces de aprender representaciones jerárquicas y abstractas de los datos, gracias al entrenamiento sobre grandes volúmenes de datos y al uso de recursos computacionales avanzados [LBH15].

En la Figura 2.1 se ilustra la relación entre la IA, el *Aprendizaje Automático* (ML) y el *Aprendizaje Profundo* (DL). La IA, en el área azul más grande, es el campo general que busca que las máquinas simulen la inteligencia humana. Dentro de este campo, el ML, mostrado en naranja, es una técnica más específica que permite a las máquinas aprender de los datos para hacer predicciones o tomar decisiones. Por último, el DL, es un campo especializado dentro del ML, que utiliza redes neuronales profundas para procesar información compleja, como imágenes o lenguaje.

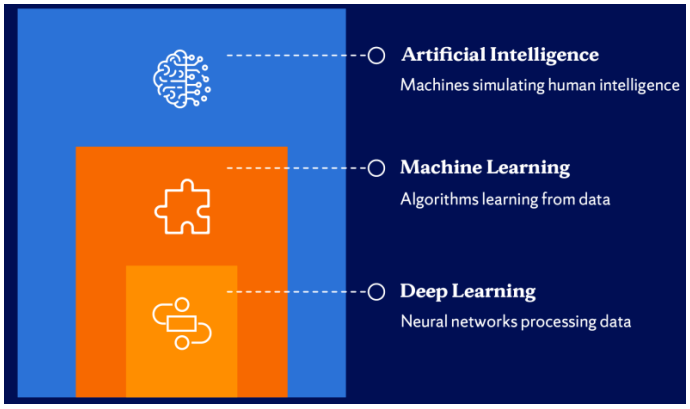


Figura 2.1: Inteligencia Artificial, Machine Learning y Deep Learning. [Syr25]

Hoy en día, la IA tiene una gran variedad de aplicaciones en diversas áreas, incluyendo la clasificación de imágenes, el reconocimiento de voz, robótica, las finanzas, los vehículos autónomos, el diagnóstico médico, los sistemas de recomendación, la visión por computador, la ciberseguridad y los videojuegos, entre otros [ZBW20]. También está transformando la forma en que nos cominicamos e interactuamos con la tecnología, accedemos a la información, trabajamos y tomamos decisiones.

2.2. Modelos Generativos

Los modelos generativos son modelos de aprendizaje automático que tienen la capacidad de aprender la distribución de probabilidad de los datos de entrada para generar nuevas muestras sintéticas similares a las reales. A diferencia de los modelos discriminativos, que aprenden la relación entre características y etiquetas para predecir o clasificar (aprendiendo P(Y—X)), los modelos generativos capturan la distribución

completa de los datos (aprendiendo P(X) o P(X, Y)).

Estos modelos se usan ampliamente para aplicaciones en síntesis de datos, como generación de imágenes, audio y texto, y aumento de datos. Entre los modelos generativos más populares se encuentran las redes generativas adversarias (GAN), los *Codificadores Automáticos Variacionales* (VAE), los modelos de difusión y los *Modelos Autorregresivos* (AR) [IBM24].

2.2.1. Redes Generativas Adversarias

Las Redes Generativas Adversarias, propuestas en 2014, son modelos compuestos por dos redes neuronales, consiste en un generador y un discriminador que compiten. El generador crea datos sintéticos, mientras que el discriminador los clasifica como reales o falsos. El objetivo es que el generador mejore hasta que tenga la capacidad de producir muestras que engañen al discriminador y así mismo el discriminador mejore su capacidad de distinguir los datos reales o sintéticos. Esta dinámica se describe por la función de pérdida minimax:

$$\min_G \max_D \; \mathbb{E}_{x\sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z\sim p_z} [\log (1 - D(G(z)))] \tag{2.1}$$

Inicialmente el generador crea muestras muy fáciles de detectar por el discriminador, pero a medida que avanza el entrenamiento, el generador produce datos cada vez más realistas que el discriminador confunde con los reales.

Las GAN tienen diversas aplicaciones en campos como la visión artificial, generación de imágenes, síntesis de audio, generación de texto condicional y datos sintéticos para entrenamiento.

2.2.2. Codificadores Automáticos Variacionales

Los Codificadores Automáticos Variacionales son modelos generativos que aprenden la distribución de probabilidad de los datos de entrada, como imágenes o audio, para generar nuevas muestras sintéticas similares a las reales. Mediante un codificador y un decodificador, los GAN generan representaciones latentes continuas en un espacio probabilístico, lo que permite modelar variaciones de los datos originales [IBM23]. En la Figura 2.2 se ilustra este proceso.

- El codificador toma los datos de entrada y los comprime en una representación más pequeña y abstracta, llamada espacio latente. En una arquitectura típica, cada capa subsiguiente del codificador contiene menos nodos que la capa anterior, por lo que a medida que los datos atraviesan cada capa del codificador, se comprimen en menos dimensiones.

- El cuello de botella (Bottleneck), es tanto la capa de salida de la red codificadora como la capa de entrada de la red decodificadora. Contiene el espacio latente: la

incrustación totalmente comprimida y de dimensiones inferiores de los datos de entrada.

- El decodificador utiliza esa representación latente para reconstruir la entrada original o generar un nuevo dato que sea similar, invirtiendo esencialmente el codificador. En una arquitectura de decodificador típica, cada capa subsiguiente contiene un número mayor de nodos activos.

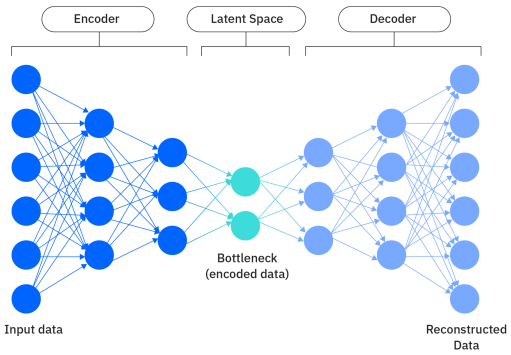


Figura 2.2: Funcionamiento de los autocodificadores variacionales. [IBM24]

2.2.3. Modelos de Difusión Latente

Los *Modelos de Difusión Latente* (LDM) se basan en generar datos nuevos aprendiendo a revertir un proceso de agregación de ruido en un espacio latente comprimido, en lugar de trabajar directamente con los datos originales (por ejemplo, píxeles de una imagen o formas de onda de audio). Estos modelos utilizan un autoencoder (generalmente un VAE) para codificar los datos en un espacio latente de menor dimensión, donde se aplica el proceso de difusión. Luego, un modelo de difusión entrenado en este espacio latente reconstruye los datos a través de un proceso inverso de eliminación de ruido, y un decodificador los convierte de vuelta al formato original [IBM24].

El funcionamiento de un LDM se basa en tres componentes principales:

1. **Codificación en el espacio latente:** Se usa un VAE para comprimir los datos de entrada en un espacio latente de menor dimensión, capturando las características semánticas esenciales del dato original.

2. **Difusión en el espacio latente:** Una vez obtenida la representación latente, se añade ruido gaussiano de manera incremental durante múltiples pasos, siguiendo una cadena de Markov. Durante el entrenamiento, una red neuronal (generalmente una arquitectura U-Net o transformer) aprende a invertir este proceso y a predecir la versión menos ruidosa en cada paso.

3. **Decodificación:** Una vez se ha completado el proceso de difusión inversa, que parte de ruido aleatorio y recupera una muestra latente coherente, se utiliza el

decodificador del VAE para reconstruir el dato en el espacio original (imagen, audio, texto, etc.). El resultado es una muestra sintética de alta calidad y fidelidad, generada a partir de ruido, pero estructurada semánticamente por la información aprendida durante el entrenamiento.

El principal beneficio de los LDM radica en que trabajan en el espacio latente, lo cual ayuda a reducir significativamente la complejidad computacional, lo que permite entrenar y ejecutar modelos generativos de difusión con muchos menos recursos en comparación con trabajar directamente sobre los datos originales [Ser22]. Además, el uso de representaciones latentes facilita la incorporación de información condicional, como texto, en el caso de los modelos de TTA.

2.2.4. Modelos Autoregresivos

Los modelos AR que tienen la capacidad de predecir el siguiente elemento en una secuencia basándose en los elementos previos. Estos modelos son particularmente útiles en la generación de secuencias, como audio o texto, donde generar una voz con un sonido natural requiere predecir la siguiente muestra de audio a partir de las muestras de audio previas, permitiendo producir resultados naturales y coherentes. La idea central detrás de un modelo autorregresivo es modelar la probabilidad conjunta de una secuencia de datos

$$\mathbf{x} = (x_1, x_2, \ldots, x_T)$$

como el producto de probabilidades condicionales:

$$P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1}) \tag{2.2}$$

Esto permite que el modelo aprenda dependencias complejas en los datos, ya que cada predicción se basa directamente en el historial previo de la secuencia.

En el contexto del aprendizaje profundo, los modelos AR pueden implementarse mediante *Redes Neuronales Recurrentes* (RNN), transformers o *Redes Neuronales Convolucionales* (CNN), como en el caso de WaveNet.

El modelo autorregresivo tiene dos componentes principales: el codificador y el decodificador. [Spe22]

- Codificador: Transforma los datos de entrada (como formas de onda o espectrogramas) en una representación intermedia (latente). Esta representación captura características esenciales, como patrones acústicos o estructuras lingüísticas
- Decodificador: Genera la salida (por ejemplo, una muestra de audio) paso a paso, utilizando cada predicción previa como contexto para la siguiente, lo que permite una generación secuencial y coherente.

2.3. Procesamiento de Señales de Audio

El procesamiento de señales de audio es una tarea esencial en los modelos de TTS y TTA, ya que permite transformar las señales acústicas en formatos que pueden ser entendidos, analizados o generados por modelos de inteligencia artificial.

Antes de que los modelos puedan generar o interpretar audio, es necesario transformar la señal de audio en representaciones más manejables y entendibles por los modelos, como espectrogramas o coeficientes cepstrales. Estas representaciones permiten que se capturen patrones relevantes del habla o del entorno sonoro, facilitando la generación de audio de alta calidad, con buena naturalidad e inteligibilidad.

Además, el procesamiento de audio incluye una serie de técnicas de preprocesamiento, como la normalización de la amplitud, la eliminación de ruido o la tokenización, que son clave para asegurar que el modelo trabaje con datos consistentes y representativos. Estas técnicas tienen un impacto directo en el rendimiento de los modelos, ya que influyen en la calidad del audio generado.

En las siguientes secciones se describen las representaciones más comunes del audio, así como las principales técnicas de preprocesamiento utilizadas en los modelos TTS y TTA.

2.3.1. Representaciones de Audio

Las señales de audio se pueden representar de múltiples formas para su procesamiento. Las representaciones de audio son transformaciones de la señal original que facilitan su análisis, procesamiento y generación. Estas pueden clasificarse en representaciones tradicionales, basadas en transformaciones matemáticas, y representaciones modernas, aprendidas por modelos de inteligencia artificial. A continuación, se describen algunas de las principales representaciones de audio [Zil25].

2.3.1.1. Representaciones Tradicionales

- **Forma de onda (Waveform)**: La forma de onda es la representación más básica de una señal de audio en el dominio del tiempo. Es una representación directa de la señal de audio, mostrando la amplitud en función del tiempo. Esta representación es útil para análisis de amplitud, detección de silencios o procesamiento de efectos. Sin embargo, no descompone las frecuencias, lo que limita su uso en análisis espectrales complejos. En la Figura 2.3 se muestra un ejemplo.

- **Espectrograma de Mel (Mel-Spectrogram)**: El espectrograma de Mel es una representación en el dominio tiempo-frecuencia que combina la *Transformada de Fourier de Corto Tiempo* (STFT)con una escala perceptual de frecuencia llamada escala Mel, que aproxima la respuesta no lineal del oído humano a diferentes frecuencias. El resultado es una imagen (frecuencia Mel vs. tiempo) donde cada píxel indica energía acústica. Es la representación estándar en modelos TTS modernos.

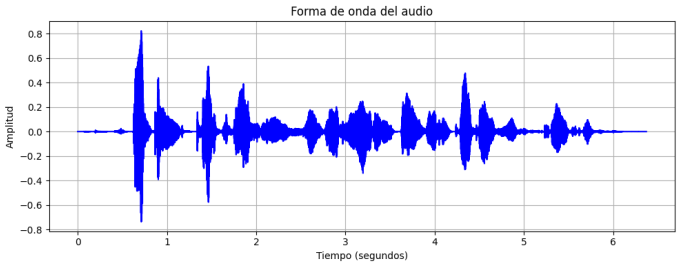


Figura 2.3: Forma de onda de un audio del conjunto de datos de Common Voice.

Los espectrogramas proporcionan una compresión útil de la información espectral, facilitando el entrenamiento. En la Figura 2.4 se muestra un ejemplo.

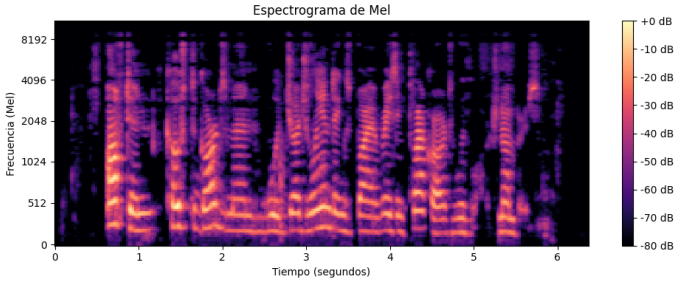


Figura 2.4: Espectograma de Mel de un audio del conjunto de datos de Common Voice.

- **Coeficientes Cepstrales (CC)**: Los *Coeficientes Cepstrales* (CC) son el resultado de aplicar una transformada inversa de Fourier al logaritmo del espectro de potencia:

$$\text{Cepstrum}(x) = \mathcal{F}^{-1}\left(\log|\mathcal{F}(x)|^2\right) \tag{2.3}$$

Esta representación separa la envolvente del espectro (información vocal) y el contenido armónico (información de tono), siendo útil en reconocimiento de voz y análisis de timbre. En la Figura 2.5 se muestra un ejemplo.

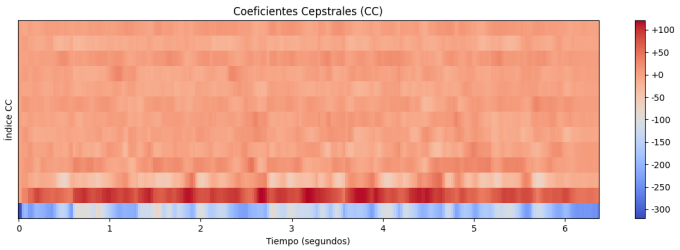


Figura 2.5: Representación cepstral de un audio del conjunto de datos de Common Voice.

- **Coeficientes Cepstrales de Frecuencia Mel (MFCCs)**: Los *Coeficientes Cepstrales de Frecuencia Mel* (MFCCs) combinan la transformación cepstral con la escala de Mel para capturar características del habla que se alinean con la percepción humana. Son una de las representaciones más populares en el procesamiento del

habla y audio. En la Figura 2.6 se muestra un ejemplo. Los pasos para calcular estos coeficientes son:

- Aplicar la STFT para dividir la señal en marcos temporales y obtener el espectro de frecuencias.
- Calcular el espectrograma de Mel, mapeando las frecuencias a la escala Mel y obteniendo las energías de las bandas.
- Aplicar la *Transformada Discreta del Coseno* (DCT) para obtener un conjunto reducido de coeficientes (típicamente 13-20 por marco).

$$MFCC_n = \sum_{k=1}^{K} \log(S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \tag{2.4}$$

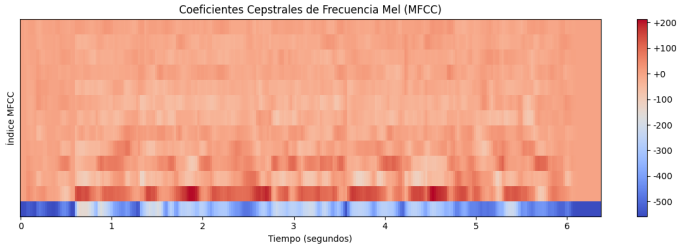


Figura 2.6: Espectrograma de los Coeficientes Cepstrales de Frecuencia Mel (MFCC) de un audio del conjunto de datos Common Voice.

2.3.1.2. Representaciones Modernas

- **Representaciones latentes (Latent representations)**: Una representación latente es una codificación comprimida de la señal de audio aprendida por una red neuronal. Modelos como autoencoders, VAE o modelos de difusión, aprenden a representar una señal en un espacio de menor dimensión (espacio latente) donde se preservan características relevantes. Estos vectores latentes comprimidos que pueden usarse como características acústicas. En la Figura 2.7 se muestra un ejemplo de la representación latente de un audio en inglés del conjunto de datos Common Voice, obtenida mediante un autoencoder. Esta representación es un vector comprimido en un espacio de 32 dimensiones que captura características acústicas relevantes de la señal de audio. En la Figura 2.8 se presentan el espectrograma de Mel original y el espectrograma de Mel reconstruido a partir de la representación latente, mostrando cómo el autoencoder preserva las características principales del audio en el proceso de compresión y reconstrucción.
- **Embeddings acústicos (Audio embeddings)**: Los embeddings acústicos son vectores numéricos que resumen las características clave de una señal de audio, generados por modelos preentrenados de aprendizaje profundo como Wav2Vec 2.0 o CLAP. Estos vectores capturan información semántica, identidad del hablante, emoción, tono, timbre, etc. aprendidos a partir de datos.

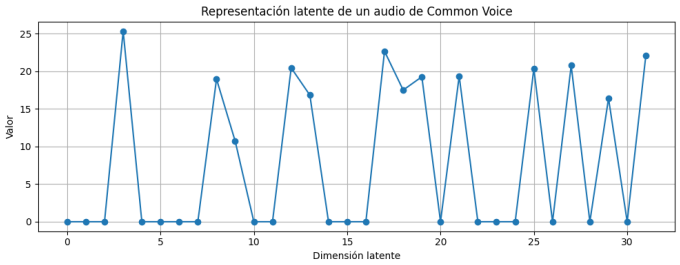


Figura 2.7: Representación latente de un audio en inglés del conjunto de datos Common Voice.

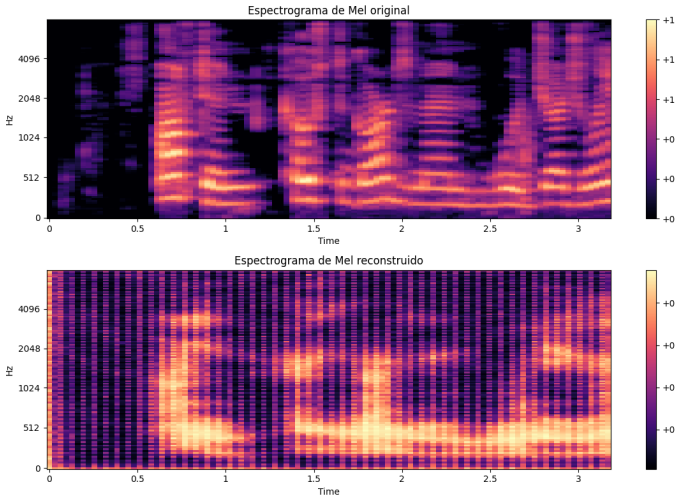


Figura 2.8: Espectrograma de Mel original y Espectrograma de Mel reconstruido a partir de la representación latente.

En la Figura 2.9 se muestra un ejemplo de los embeddings acústicos promedio de un audio en inglés del conjunto de datos Common Voice, obtenidos mediante el modelo Wav2Vec 2.0 (`facebook/wav2vec2-large-xlsr-53`). Este vector de 1024 dimensiones, calculado promediando los embeddings contextuales a lo largo del tiempo, representa las características acústicas comprimidas del audio, incluyendo tono, timbre y contexto fonético. Complementariamente, la figura 2.10 presenta los embeddings completos como un mapa de calor, mostrando la evolución de las 1024 dimensiones a lo largo de los frames temporales del audio

2.3.2. Técnicas de Preprocesamiento

El preprocesamiento prepara los datos de audio y texto para su uso en modelos como TTS y TTA, mejorando la calidad y consistencia de los resultados.

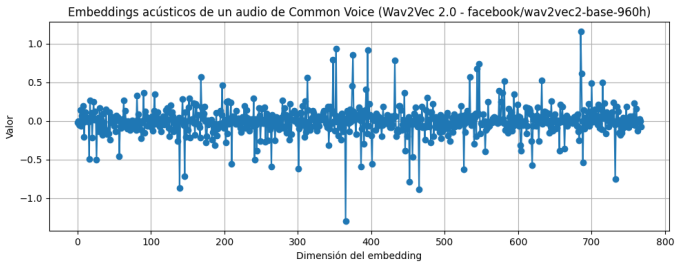


Figura 2.9: Embeddings acústicos promedio de una grabación en inglés del conjunto de datos Common Voice, generados con el modelo Wav2Vec 2.0 (`facebook/wav2vec2-large-xlsr-53`)

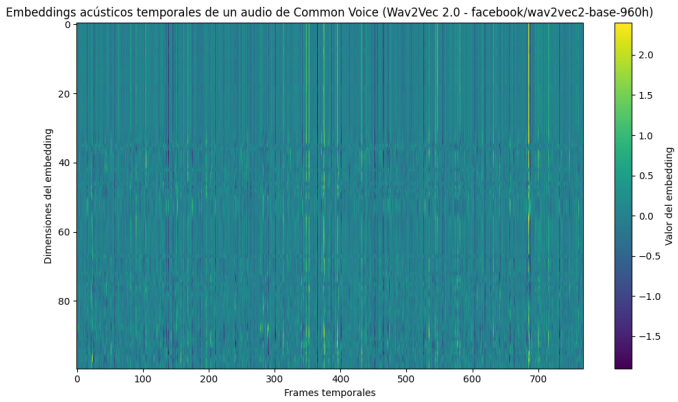


Figura 2.10: Embeddings acústicos temporales de una grabación en inglés del conjunto de datos Common Voice, generados con el modelo Wav2Vec 2.0 (`facebook/wav2vec2-large-xlsr-53`), visualizados como un mapa de calor.

2.3.2.1. Preprocesamiento de Audio

Consiste en transformar señales de audio crudas en un formato adecuado para su análisis o modelado, mejorando su calidad y compatibilidad con algoritmos de aprendizaje automático. A continuación, se describen las técnicas más relevantes [Zil25].

- **Normalización de la amplitud**: Ajustar el volumen de una señal para alcanzar un valor objetivo sin alterar su rango dinámico. De esta forma se logra que se encuentre dentro de un rango específico, evitando distorsiones o desbalances. Se utiliza en diversas aplicaciones, como la preparación de archivos para streaming, la mejora de la coherencia entre múltiples pistas y la optimización del volumen en la edición de audio.
- **Eliminación de ruido (Noise reduction)**: Aplicar métodos y algoritmos para reducir o remover el ruido, eliminar sonidos de fondo constantes o interferencias. Esto mejora la relación señal/ruido, hace más limpia la voz y mejora la claridad del audio. En conjuntos de datos grandes se suele usar bibliotecas especializadas (noisereduce en Python) para suavizar o eliminar zumbidos y estática.
- **Reesampleo (Resampling)**: Ajustar la tasa de muestreo para segurar que

todas las muestras de audio tengan la misma frecuencia y así cumplir con los requisitos del modelo. La mayoría de los modelos preentrenados en audio usan 16 kHz. Así, si los datos originales están a 8 kHz o a 48 kHz, se los remuestrea a la tasa estándar de entrenamiento. Para esto se utilizan bibliotecas como librosa en Python.

- **Segmentación (Chunking)**: Dividir el audio en segmentos más cortos (por ejemplo, clips de 20 segundos), para facilitar el procesamiento, el entrenamiento y evitar problemas de memoria en modelos con rrstricciones de longitud.

2.3.2.2. Preprocesamiento de Texto

El preprocesamiento de texto implica transformar datos textuales crudos en un formato estructurado que facilite su análisis. A continuación, se describen las técnicas más relevantes, basadas en su función y aplicación práctica [Pab24].

- **Tokenización**: Es el primer paso en el preprocesamiento de texto y consiste en dividir un texto en unidades más pequeñas, denominadas *tokens*, que pueden ser palabras, frases, o incluso signos de puntuación. Este proceso permite descomponer textos complejos en elementos manejables para su análisis posterior. Por ejemplo, la oración "La inteligencia artificial transforma el análisis de datos" se tokeniza en:

$$\text{Tokens} = [\text{La}, \text{inteligencia}, \text{artificial}, \text{transforma}, \text{el}, \text{análisis}, \text{de}, \text{datos}]$$

  La tokenización puede realizarse a nivel de palabras (como en el ejemplo anterior) o de oraciones, dependiendo de la tarea. Herramientas como NLTK o spaCy en Python facilitan este proceso, permitiendo personalizar la inclusión de signos de puntuación o el manejo de mayúsculas openwebinars2024python. La tokenización es crucial para tareas como el etiquetado de partes del discurso (*Part-of-Speech Tagging*) o la construcción de modelos de *Bag of Words*. La tokenización puede realizarse a nivel de palabras (como en el ejemplo anterior) o de oraciones, dependiendo de la tarea. Herramientas como NLTK o spaCy en Python facilitan este proceso, permitiendo personalizar la inclusión de signos de puntuación o el manejo de mayúsculas.
- **Lematización**: Reduce las palabras a su forma base o *lema*, que es la forma estándar que aparece en un diccionario, considerando el contexto gramatical. Por ejemplo, las palabras "corriendo", "corre" y "corrí" se lematizan a "correr". Este proceso utiliza diccionarios y reglas gramaticales para garantizar que la reducción preserve el significado. La lematización es más precisa que el *stemming* porque tiene en cuenta el contexto, lo que la hace ideal para tareas que requieren comprensión semántica, como el análisis de sentimientos o la traducción automática. Un ejemplo práctico sería:

$$\text{Lematización: } [\text{cantando}, \text{cantó}, \text{cantará}] \rightarrow [\text{cantar}, \text{cantar}, \text{cantar}]$$

En Python, bibliotecas como spaCy ofrecen herramientas robustas para la lematización en múltiples idiomas, incluyendo el español.

- **Stemming**: Es otra técnica de normalización que reduce las palabras a su raíz o forma base, eliminando sufijos o prefijos, pero sin considerar el contexto gramatical. Por ejemplo, "cantando" y "cantador" se reducen a "cant". A diferencia de la lematización, el *stemming* es más simple y rápido, pero menos preciso, ya que puede generar raíces no válidas (por ejemplo, "universidad" y "universal" podrían reducirse a "univ"). Un ejemplo sería:

  $$\text{Stemming: } [\text{cantando}, \text{cantador}] \rightarrow [\text{cant}, \text{cant}]$$

  El *stemming* es útil en aplicaciones donde la velocidad es prioritaria, como la búsqueda de información o la indexación de documentos. Algoritmos como el de Porter o Snowball, disponibles en NLTK, son comúnmente utilizados para este propósito.

- **Eliminación de Palabras Vacías**: La eliminación de palabras vacías (*stop words*) consiste en filtrar palabras comunes que no aportan significado relevante al análisis, como "el", "de", "y" o "es" en español. Estas palabras son frecuentes en el lenguaje, pero suelen ser irrelevantes para tareas como la clasificación de textos o el análisis de temas. La eliminación de *stop words* reduce el ruido en los datos y mejora la eficiencia computacional al disminuir el tamaño del vocabulario. Por ejemplo:

  Texto original: La inteligencia artificial transforma el análisis de datos

  Tras eliminar stop words: [inteligencia, artificial, transforma, análisis, datos]

  Listas predefinidas de *stop words* están disponibles en bibliotecas como NLTK o spaCy, aunque pueden personalizarse según el contexto de la aplicación.

2.4. Texto a Voz

Los sistemas de TTS, o conversión de texto a voz, son una tecnología con la capacidad de transformar el texto escrito en audio hablado. Este proceso no solo busca sintetizar palabras, sino también lograr un habla que sea natural y comprensible, imitando el tono, la entonación y otras características prosódicas, con el propósito de simular la comunicación verbal humana de manera realista.

La importancia de la conversión de texto a voz ha ido aumentando crecientemente debido a sus aplicaciones en diversos campos [MRVPK23], como la accesibilidad para personas con discapacidades visuales, asistentes virtuales como Google Assistant, Siri y Cortana, sistemas de navegación, e-learning, audilibros y herramientas de servicio al cliente [WCM+21, KKS23, AR24]. Las metodologías implementadas para generar voz a partir de texto han ido evolucionando a través del tiempo, con enfoques tradicionales basados

en reglas, modelos concatenativos y paramétricos como la *Síntesis de Voz Concatenativa* (CSS), *Síntesis de Voz Paramétrica* (PSS), *Síntesis de Voz Paramétrica Estadística* (SPSS) hasta sistemas modernos basados en aprendizaje profundo [MRVPK23, NM21] como las *Máquinas de Boltzmann Restringidas* (RBM) [LDY13], las CNN [TUA18], las RNN, los *Modelos de Memoria a Largo Plazo* (LSTM) y las GAN. ofreciendo resultados que combinan la naturalidad, flexibilidad, comprensibilidad y la preferencia del habla sintética [KKS23, KPK+23]. En la Figura 2.11 se presenta la ilustración del flujo del proceso de la arquitectura básica de un modelo de TTS.

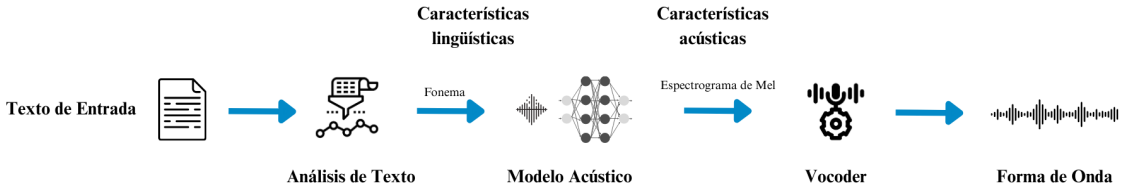


Figura 2.11: Estructura General de los Modelos de TTS.

El funcionamiento de los sistemas TTS modernos constan generalmente de 3 elementos principales [KPK+23]:

1. **Análisis de texto:** Este módulo procesa el texto de entrada para extraer las características lingüísticas, las cuales contienen información relevante sobre la pronunciación y la prosodia [Has23]. Modelos como Char2Wav [SMK+17] y DeepVoice [GAD+17a] se basan unicamente en las redes neuronales para realizar la conversión de las características lingüísticas [TQSL21].

2. **Modelo acústico:** Convierte las características lingüísticas obtenidas en el análisis de texto en representaciones acústicas como mel-espectrogramas, los cuales codifican las frecuencias y amplitudes de los sonidos del habla a lo largo del tiempo. Modelos como Tacotron [WSRS+17], FastSpeech [RRT+19] y FastPitch [Łań21] utilizan redes neuronales para realizar esta tarea de manera eficiente, logrando una alineación implícita entre las características lingüísticas y acústicas, lo que mejora la naturalidad y claridad del habla sintetizada [CLL21].

3. **Vocoder:** Convierte las representaciones acústicas intermedias, como los mel-espectogramas en formas de onda de audio crudo, generando el sonido final que escuchará el usuario. Los vocoders se clasifican en dos categorías principales: los tradicionales basados en el procesamiento estadístico de señales (SPSS) y los modernos basados en redes neuronales [TQSL21]. Ejemplos de vocoders tradicionales son STRAIGHT y WORLD, que toman características acústicas como entrada [KS23]. Por otro lado, los vocoders basados en redes neuronales, como WaveNet, WaveGlow y MelGAN, ofrecen una mayor calidad de síntesis al procesar tanto características acústicas como lingüísticas, dependiendo del modelo [KPK+23].

Los modelos basados en aprendizaje profundo pueden ser clasificados en regresivos y no autoregresivos [KPK+23, TSİ+23]. Los modelos autoregresivos, como Wavenet [NM21],

generan el discurso de manera secuencial, donde cada paso depende del anterior, lo que puede llevar a problemas de latencia y sufrir tiempos de procesamiento más lentos [TSİ+23]. En contraste, los modelos no autoregresivos, como FastSpeech, producen las salidas en paralelo, mejorando significativamente la velocidad de síntesis [KPK+23, KS23]. Sin embargo, estos modelos requieren métodos adicionales para alinear las entradas y salidas, como el uso de reguladores de longitud o alineadores externos [KPK+23].

A pesar de los avances logrados, los sistemas TTS enfrentan varios desafíos [KKS23]. Uno de los principales es lograr mejorar la naturalidad y la inteligibilidad del habla sintetizada para conseguir que el audio sintético generado sea indistinguible del humano, lo cual implica desarrollar algoritmos y técnicas robuztas que tengan la capacidad de capturar emociones, prosodia compleja, variaciones en el timbre de voz y la preservación de las características del hablante [MRVPK23, SNT24]. Además, el entrenamiento de modelos neuronales modernos requiere grandes cantidades de datos etiquetados y recursos computacionales significativos, lo que limita su aplicabilidad en idiomas con pocos recursos como los idiomas indios [KPK+23, AR24].

2.5. Texto a Audio

La generación de TTA es una técnica de inteligencia artificial que permite la síntesis de audio a partir de descripciones textuales detalladas. A diferencia de los modelos de TTS que se enfocan exclusivamente en la generación de la voz humana a partir de texto, centrándose principalmente en aspectos como la naturalidad, la entonación y la prosodia [KKS23], los modelos de TTA no se limitan únicamente a generar habla, sino que buscan integrar diferentes comportamientos acústicos y contextos sonoros a partir de las descripciones textuales proporcionadas en la entrada [LCY+23], generando audios con efectos de sonido (como “el timbre de una puerta”), música (como “una melodía de piano”), ambientes acústicos (como “una tormenta con truenos y lluvia intensa”) o incluso secuencias de eventos acústicos complejos (como .[el] ladrido de un perro seguido del ruido de una sirena de bomberos”) [HJL+24].

Estos modelos han avanzado rápidamente y han mostrado su potencial en diversas áreas como el entretenimiento, la educación, la producción cinematográfica y la realidad virtual [DCL+24]. En la producción audiovisual, los modelos TTA permiten generar efectos de sonido personalizados para películas, videojuegos y realidad virtual [YYW+23]. También se utilizan en la creación de experiencias inmersivas en entornos de realidad aumentada y simulaciones acústicas [LCY+23]. La evolución de los modelos TTA ha dado lugar a enfoques más sofisticados que superan las limitaciones de las primeras propuestas, como las que empleaban etiquetas one-hot [XDGL24]. Los avances más recientes han mejorado la calidad y naturalidad del audio generado, integrando nuevas técnicas como los modelos de lenguaje y los modelos de difusión latente [XDGL24]. Modelos recientes, como AudioLDM y Auffusion, han demostrado la efectividad de los modelos LDM en la mejora de la calidad del audio generado, al tiempo que reducen los requisitos de datos emparejados [XDGL24,

LCY+23]. Además de los avances en arquitectura, otro desafío importante en el desarrollo de modelos TTA ha sido la escasez de datos de entrenamiento de alta calidad. Para abordar este problema, se han propuesto métodos innovadores de aumento de datos, como la mezcla basada en niveles de presión de audio. Este enfoque permite generar conjuntos de datos más equilibrados, superando las limitaciones de combinaciones aleatorias utilizadas en sistemas anteriores [GMMP23]. En la Figura 2.12 se presenta un ejemplo de la arquitectura de los sistemas TTA.

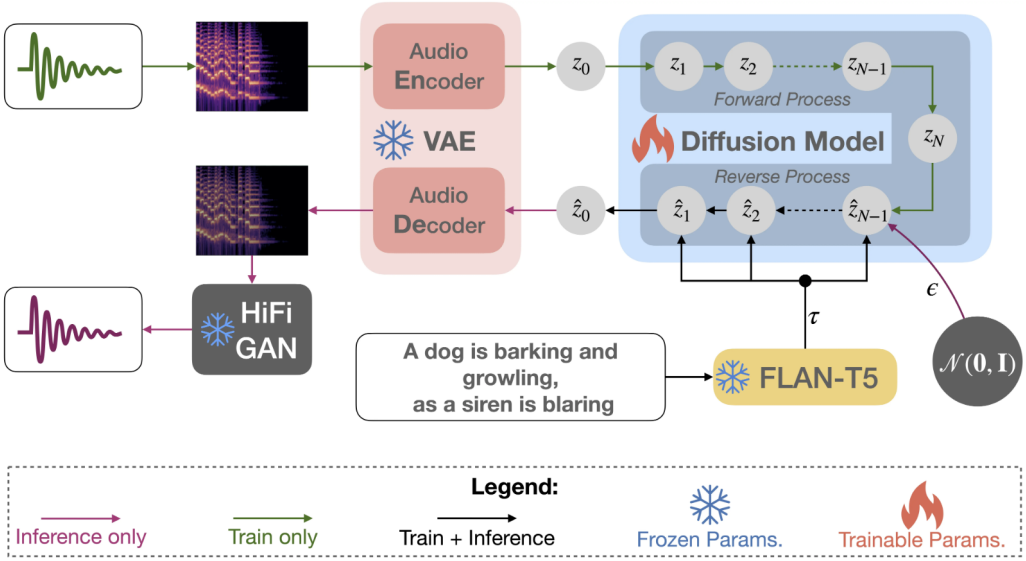


Figura 2.12: Estructura General de los Modelos TTA [Sum24].

Los modelos avanzados de TTA logran resultados notables gracias a su capacidad para alinear de manera efectiva el texto y el audio mediante representaciones latentes que capturan las relaciones entre ambos elementos [HHY+23]. Tecnologías como CLAP y Flan-T5, que aprovechan preentrenamientos en grandes conjuntos de datos, han demostrado ser particularmente útiles al mapear descripciones textuales en características acústicas coherentes [YLL+24, WTT24]. Por ejemplo, Auffusion adapta metodologías desarrolladas para la generación de *Texto a Imagen* (T2I) al contexto de TTA, logrando una alineación precisa entre texto y audio que permite generar sonidos estilizados y personalizados con gran fidelidad [XDGL24].

Los modelos de TTA enfrentan varios desafíos inherentes como el sesgo en los datos y falta de diversidad [KSP+22]. Los modelos entrenados en conjuntos de datos desequilibrados tienden a generar correctamente sonidos comunes, pero fallan en representar clases de audio menos frecuentes. Este problema, conocido como çola larga", ha sido abordado mediante técnicas de recuperación aumentada, como Re-AudioLDM [YLL+24]. Otro de los desafíos a los que se enfrentan los modelos de TTA es modelar las relaciones entre los eventos de audio descritos en el texto de entrada, como la secuencia precisa de sonidos en descripciones complejas. Modelos recientes como RiTTA [HJL+24] han intentado abordar esta limitación, utilizando enfoques específicos para mejorar la capacidad de los modelos TTA existentes para modelar las relaciones de eventos de audio.

2.6. Síntesis del Capítulo

Este capítulo ha establecido las bases teóricas esenciales para la generación de audio sintético multilingüe y multiacento, reforzando el papel de la IA como tecnología clave en el avance de la síntesis de audio. Los avances en aprendizaje profundo, particularmente en modelos generativos como redes adversarias, codificadores variacionales, modelos de difusión y autorregresivos, han transformado la capacidad de generar audio realista, sentando un precedente para abordar desafíos como la naturalidad y la personalización. El análisis del procesamiento de señales de audio destaca la importancia de representaciones como espectrogramas de Mel y embeddings acústicos, junto con técnicas de preprocesamiento, para garantizar la calidad y consistencia de los datos en modelos TTS y TTA. Asimismo, se reconoce que los sistemas TTS han alcanzado niveles significativos de naturalidad, pero persisten limitaciones en la captura de prosodia compleja y la aplicabilidad a idiomas con pocos recursos. Por su parte, los modelos TTA amplían el alcance de los modelos TTS al integrar contextos acústicos, aunque enfrentan retos en la alineación texto-audio y la diversidad de datos. Con estos análisis se evidencia la necesidad de un enfoque integrado que combine las fortalezas de ambas técnicas para superar las limitaciones actuales en la generación de audio sintético.

Capítulo 3

Estado del Arte

En este capítulo se presenta un análisis detallado del estado del arte en la generación de audio sintético, centrado en los avances y desafíos de los modelos de TTS, TTA y aquellos que combinan ambas capacidades. En las Secciones 3.1, 3.2 y 3.3, se examinan detalladamente las arquitecturas, enfoques y limitaciones de modelos representativos, incluyendo aquellos basados en Transformers, difusión, redes neuronales y modelado de lenguaje. Además, se incluyen tablas comparativas, la Tabla 3.1, que resume las características técnicas de los modelos analizados, como sus arquitecturas, vocoders y *Puntuación de opinión media* (MOS) y la Tabla 3.2 que compara los conjuntos de datos utilizados.

La generación de audio sintético se ha convertido en un área clave de investigación en el campo del aprendizaje profundo, impulsada por la creciente demanda de aplicaciones que requieren voces realistas, sonidos personalizados y narraciones automatizadas [BTS+24]. Este campo abarca una variedad de técnicas y modelos diseñados para producir diferentes tipos de audio, desde habla humana hasta efectos sonoros y composiciones musicales. A lo largo de los años, los avances en redes neuronales profundas, las arquitecturas basadas en transformadores y las GAN, han permitido un progreso significativo en la calidad, flexibilidad y realismo del audio generado [BH24]. Entre los enfoques principales, se destacan los modelos de TTS [KVB+23, Bet23, CWS+22], que convierten texto en habla realista; los modelos de TTA [KSP+22, YYW+23, XDGL24], que amplían el alcance generando no sólo voz, sino también música, efectos de sonido y ambientes acústicos a partir de descripciones textuales. Además, han surgido sistemas avanzados que combinan estas técnicas para generar audios sintéticos con mayor versatilidad. Estos modelos no solo pueden crear audios a partir de descripciones textuales, sino también replicar características prosódicas o ambientales de un audio de referencia, lo que permite una síntesis más personalizada y realista. [LYNC24, CDG+24, CWS+22].

3.1. Modelos de Texto a Voz

3.1.1. Modelos Basados en Transformers

[KVB+23] SPEAR-TTS es un sistema de TTS multi-hablante que se puede entrenar con una supervisión mínima. Este modelo se basa en tres componentes principales: generación de texto semántico, predicción de tokens acústicos y síntesis de formas de onda. Utiliza un enfoque basado en Transformers para procesar los datos de texto y audio, combinando modelos previos de codificación jerárquica como w2v-BERT y sistemas de vocoder avanzados como HiFi-GAN. Su arquitectura permite generalizar a nuevos locutores. Utiliza un enfoque de ”prompting”para controlar la identidad del hablante, lo que permite sintetizar voces nuevas utilizando solo una muestra corta de voz de 3 segundos. Para el entrenamiento de este modelo se utilizaron los dataset: LibriLight, LibriTTS y LJSpeech, que incluyen grabaciones de voz etiquetadas y no etiquetadas. El modelo ofrece una síntesis de voz de alta fidelidad, logrando una calidad cercana a la voz humana incluso con supervisión limitada y reduce significativamente la dependencia de datos de texto y audio emparejados, pero requiere una cantidad considerable de datos, del orden de miles de horas. Esto puede ser una limitación para idiomas con pocos recursos.

3.1.2. Modelos Basados en Difusión

Los autores de [LGS+] proponen PromptTTS, un modelo de TTS basado en indicaciones de texto que tiene como objetivo mejorar la síntesis de voz al permitir que las características del hablante, como el tono, el timbre y el estilo, sean especificadas a través del texto de entrada, facilitando la personalización de la voz sin depender de muestras de audios de referencia. El modelo introduce dos componentes principales para mejorar la generación de voz: una Red de Variación para proporcionar información de la variabilidad de la voz que no está capturada por indicaciones de texto, y un canal de generación de indicaciones para utilizar *Modelos de Lenguaje Grandes* (LLM) para componer indicaciones de texto de alta calidad. La red de variación predice la representación extraída del habla de referencia (que contiene información completa sobre la variabilidad de la voz) basándose en la representación del mensaje de texto. Para el proceso de generación de indicaciones, genera indicaciones de texto para el habla con un modelo de comprensión del lenguaje del habla para reconocer atributos de voz (por ejemplo, género, velocidad) del habla y un modelo de lenguaje grande para formular indicaciones de texto basadas en los resultados del reconocimiento. El modelo fue entrenado con diversos conjuntos de datos como AudioSet, BBC sound effects, Audiostock, AudioCaps-train, ESC-50, FSD50K, Free To Use Sounds, Sonniss Game Effects. El modelo fue evaluado con la métrica objetiva MOS, obtetiendo ua puntuación de 3,88. Una de las limitaciones que presenta el modelo es que la calidad de la síntesis puede depender de la especificidad y claridad de las descripciones textuales proporcionadas.

3.1.3. Modelos con Arquitectura Codificador-Decodificador

En [CDG+24] proponen un modelo de TTS de disparo cero que destaca por su capacidad multilingüe, abarcando 16 idiomas incluyendo aquellos con recursos bajos o medios. El modelo XTTS se basa en el modelo Tortoise, pero incorpora modificaciones clave para mejorar el entrenamiento multilingüe, la clonación de voz y la velocidad de entrenamiento e inferencia. El modelo utiliza un codificador GPT-2 con 443 millones de parámetros para procesar las entradas de texto y predecir los códigos de audio del VQ-VAE. XTTS Implementa un sistema de aprendizaje zero-shot que mapea el texto de entrada a representaciones acústicas, incluso para idiomas no vistos durante el entrenamiento e incluye mecanismos para manejar variaciones lingüísticas y prosódicas a través de embeddings que codifican la estructura fonética y el contexto lingüístico. Los datasets utilizados para entrenar el modelo son: LibriTTS, LibriLight, Common Voice. XTTS es capaz de sintetizar voz en 16 idiomas, incluyendo aquellos con recursos bajos o medios, lo que lo diferencia de la mayoría de los modelos ZS-TTS que se centran en un solo idioma o en unos pocos idiomas de altos recursos. Aunque XTTS puede ajustarse a nuevos hablantes, puede tener dificultades para clonar con precisión la voz de hablantes con características muy diferentes a las de los hablantes presentes en el conjunto de datos de entrenamiento.

Se propone un modelo diseñado para ofrecer síntesis de voz multi-locutor y conversión de voz en escenarios de cero disparos (zero-shot) [CWS+22]. Este enfoque permite generar voces completamente nuevas o convertir una voz a otra sin requerir datos de entrenamiento adicionales para las voces objetivo. Una de las principales novedades de YourTTS es el uso de texto sin procesar en lugar de fonemas como entrada, lo que facilita su aplicación a idiomas con recursos limitados. Utiliza una arquitectura de Transformer para mapear texto a espectrogramas Mel de manera eficiente, integra representaciones latentes de identidad vocal que permiten generar o convertir voces a partir de muestras de audio extremadamente cortas (tan solo 3 segundos). Además, emplea técnicas de alineación no supervisadas para garantizar una correspondencia precisa entre texto, identidad vocal y audio generado. Los datasets que se utilizaron para entrenar el modelo son VCTK, TTS-Portugiese Corpus y fr FR set del conjunto de datos M-AILABS. Se realizaron experimentos en los conjuntos de datos VCTK, LibriTTS y MLS-PT y se evaluaron las métricas SECS, MOS y Sim-MOS. En el conjunto de datos VCTK, los mejores resultados de similitud se obtuvieron con Sys.1 (monolingüe) y Sys.2 + SCL (bilingüe), alcanzando un SECS de 0,864.

3.1.4. Modelos Basados en Redes Neuronales

Los autores de [WSRS+17] propusieron Tacotron, un modelo generativo de texto a voz de extremo a extremo que sintetiza el habla directamente a partir de caracteres. El objetivo de este modelo es simplificar el proceso de síntesis de voz mediante una arquitectura unificada que permita mapear directamente texto a espectrogramas Mel, mejorando la calidad del audio generado y la facilidad de implementación. El modelo se basa en la arquitectura de secuencia a secuencia (seq2seq) con atención, donde un codificador procesa la secuencia de entrada de caracteres y un decodificador genera un espectrograma de audio. Para abordar los desafíos específicos de la síntesis de voz, Tacotron incorpora varias técnicas clave: un codificador de caracteres que utiliza una red neuronal recurrente (RNN) para capturar el contexto de los caracteres de entrada; el decodificador, también basado en RNN, utiliza un mecanismo de atención para alinear selectivamente la salida con las partes relevantes de la entrada, lo que permite una generación de voz más precisa y natural; y un vocoder de postprocesamiento para convertir el espectrograma generado en una forma de onda de audio audible, para esto utiliza el algoritmo Griffin-Lim. Tacotron, entrenado en un conjunto de datos interno de voz en inglés de un solo hablante, que contiene aproximadamente 24,6 horas de datos de voz, alcanza una MOS de 3,82 en una escala de 5 puntos, superando a un sistema paramétrico de producción en términos de naturalidad.

[SMK+17] propusieron un enfoque totalmente integrado para la síntesis de texto a voz (TTS) que simplifica el proceso de generación de audio y permite una síntesis más natural y eficiente, al eliminar la necesidad de pasos intermedios tradicionales y permitiendo la generación directa de formas de onda desde el texto de entrada. Char2Wav es un modelo de síntesis de voz de extremo a extremo que consta de dos componentes principales: un lector y un vocoder neuronal. El lector, un modelo codificador-decodificador con atención, procesa el texto o los fonemas de entrada y produce características acústicas. Estas características luego alimentan al vocoder neuronal, una extensión condicional de SampleRNN, que genera muestras de forma de onda sin procesar. En cuanto a los datasets utilizados, para la generación de voz en inglés a partir de fonemas y texto, se utilizó el dataset VCTK y para la generación de voz en español, se empleó el dataset DIMEX-100. Los resultados muestran que el modelo es capaz de sintetizar voz inteligible de alta calidad y muestra una gran robustez a errores de transcripción.

Los autores de [GAD+17b, PPG+18] presentan Deep Voice 2 y Deep Voice 3, dos sistemas neuronales de texto a voz que representan una evolución en la síntesis de voz. Ambos modelos comparten el objetivo de extender la generación de voz de alta calidad a partir de entradas de texto a múltiples locutores, pero difieren en su arquitectura y enfoque. Deep Voice 2 [GAD+17b], presenta en su arquitectura la separación del modelo de duración de fonemas y el modelo de frecuencia, lo que permite una predicción más precisa de la duración y el perfil de frecuencia de los fonemas. Además, incorpora un vocoder neuronal basado en WaveNet para la conversión de espectrograma a audio. Para la síntesis de voz multi-hablante, introduce vectores de embeddings de hablante entrenables

de baja dimensión, los cuales permiten al modelo generar diferentes voces a partir de un solo modelo, lo que reduce la necesidad de datos y esfuerzo de desarrollo para cada hablante individual. El modelo Deep Voice 2 fue entrenado con los datasets VCTK y AudioBooks. Por su parte, Deep Voice 3 [PPG+18] da un paso más allá al emplear una arquitectura totalmente convolucional basada en atención. Esta arquitectura permite un cálculo completamente paralelo, lo que se traduce en un entrenamiento significativamente más rápido en comparación con las arquitecturas recurrentes utilizadas en Deep Voice 2. También introduce mecanismos para mitigar los errores comunes en los modelos de síntesis basados en atención, como la repetición u omisión de palabras. Además, escala a conjuntos de datos de audio considerablemente más grandes, lo que le permite aprender una mayor variedad de voces y estilos de habla. El modelo fue entrenado con los datasets VCTK y LibriSpeech. En cuanto a los resultados en las MOS, Deep Voice 3 logra las puntuaciones más altas con los tres vocoders (Griffin-Lim, WORLD y WaveNet). Con el Vocoder WaveNet, Deep Voice 3 obtiene una puntuación de 3,78 mientras que Deep Voice 2 obtiene una puntuación de 2,74.

[RRT+19] proponen el modelo de síntesis de texto a voz (TTS) FastSpeech con el objetivo de abordar las limitaciones de velocidad, robustez y control en los sistemas TTS autorregressivos convencionales. La arquitectura de FastSpeech consta de una red neuronal feed-forward basada en Transformer para generar espectrogramas mel en paralelo, lo que acelera significativamente el proceso de síntesis. Para lograr la generación paralela de espectrogramas mel, utiliza un regulador de longitud que ajusta la longitud de la secuencia de fonemas de entrada para coincidir con la longitud de la secuencia de espectrogramas mel de destino. La predicción precisa de la duración asegura una alineación estricta entre los fonemas y sus correspondientes espectrogramas mel, lo que reduce los errores de omisión o repetición de palabras que pueden ocurrir en los modelos autorregresivos. Además, FastSpeech permite el control de la velocidad de la voz al ajustar la duración de los fonemas a través del regulador de longitud. También se puede controlar la prosodia al agregar pausas entre fonemas adyacentes. Los experimentos realizados en el conjunto de datos LJSpeech demuestran que FastSpeech logra una calidad de voz comparable a los modelos autorregresivos mientras acelera la generación de espectrogramas mel en 270x y la síntesis de voz completa en 38x.

FastPitch, el modelo propuesto por [Łań21], es un modelo de síntesis de texto a voz (TTS) completamente paralelo basado en FastSpeech, con el objetivo de predecir y controlar la generación de voz en contornos de frecuencia fundamental (tono). A diferencia de modelos autorregresivos como Tacotron 2, FastPitch genera espectrogramas mel en paralelo, lo que lo hace significativamente más rápido. El modelo tiene la capacidad de predecir un valor de tono para cada símbolo de entrada, lo que permite un control preciso sobre la expresividad y la prosodia del habla generada. FastPitch también puede utilizarse para generar variaciones de voz con tonos más agudos o graves, manteniendo la identidad percibida del hablante. Combinado con el vocoder WaveGlow, FastPitch puede sintetizar espectrogramas mel a una velocidad 60 veces mayor que la del tiempo real. El modelo fue entrenado con el dataset LJSpeech, que contiene aproximadamente 24 horas de habla

de un solo hablante grabada a 22 050 Hz. El modelo fue evaluado con la métrica MOS, obteniendo un puntaje de 4,080.

FAIRSEQ [WHA+21] es un popular kit de herramientas de código abierto para el modelado de secuencias basado en PyTorch, diseñado para la síntesis de voz. Permite generar formas de onda de voz con características específicas, como contenido textual, identidad del hablante y estilos de habla, mediante la implementación de modelos autorregresivos AR, como Tacotron 2 y no autorregresivos, como FastSpeech 2, junto con sus variantes multihablante. Para la conversión de espectrogramas a formas de onda, FAIRSEQ integra el vocodificador Griffin-Lim y soporta vocodificadores neuronales como WaveGlow y HiFiGAN. Una contribución clave es su pipeline de preprocesamiento de audio, que incluye eliminación de ruido de fondo, detección de actividad de voz y filtrado de outliers basado en la relación señal-ruido y la tasa de error de carácter, facilitando el entrenamiento con datos menos procesados. Los experimentos se realizaron en datasets como LJSpeech (síntesis de un solo hablante), VCTK (síntesis multihablante) y la porción en inglés de Common Voice (síntesis multihablante con datos ruidosos). Los resultados muestran que FastSpeech 2 y Transformer alcanzan un rendimiento comparable en LJSpeech, con MOS de 4.15 y 4.18 respectivamente.

3.1.5. Modelos Basados en Alineación por Refuerzo

El modelo Reinforce-aligner propuesto en [CLL21] está diseñado para abordar los desafíos de alineación en los sistemas de síntesis de texto a voz (TTS) de extremo a extremo. Este modelo introduce un enfoque basado en aprendizaje por refuerzo para mejorar la precisión y robustez de la alineación entre las secuencias de texto y las características acústicas. El modelo se basa en un agente, que es un predictor de duración de fonemas, que interactúa con el entorno, que es la red de texto a forma de onda. El agente predice la duración de cada fonema, y esta predicción se utiliza para generar la forma de onda de audio. La calidad del audio generado se evalúa mediante una función de recompensa, que se basa en la pérdida del espectrograma mel. El agente recibe retroalimentación de la recompensa y ajusta sus predicciones de duración para maximizar la recompensa. Este proceso de aprendizaje por refuerzo permite al alineador optimizar la alineación de la duración de los fonemas y mejorar la fidelidad y naturalidad del audio sintetizado. Los datasets utilizados para el entrenamiento del modelo son Korean Language (NIKL) corpus y LJSpeech. En cuanto a los resultados, el modelo fue evaluado con métricas subjetivas y obtuvo una MOS de 4,07, lo cual indica que la alineación robusta lograda por el modelo contribuye a una calidad de audio percibida positivamente.

3.2. Modelos de Texto a Audio

3.2.1. Modelos Basados en Difusión

Se propone el modelo Auffusion [XDGL24], que es un sistema de generación de texto a audio que adapta los marcos de modelos de difusión de T2I a la tarea de TTA, aprovechando las fortalezas generativas de los modelos T2I y su capacidad para una alineación intermodal precisa. El objetivo de Auffusion es abordar las limitaciones en la calidad de generación y la alineación texto-audio que se observan en los sistemas TTA existentes, especialmente cuando se manejan entradas de texto complejas. A diferencia de los modelos TTA anteriores que a menudo requieren grandes conjuntos de datos y recursos computacionales, Auffusion logra un rendimiento superior utilizando datos y recursos limitados. Auffusion se basa en un LDM preentrenado para tareas T2I, lo que le permite aprovechar la comprensión de alineación intermodal del LDM. El modelo fue entrenado con los datasets AudioCaps, WavCaps, MACS, Clotho, ESC50, UrbanSound, Musci Instruments y GTZAN. Los resultados de las evaluaciones objetivas y subjetivas demuestran que Auffusion supera los enfoques TTA anteriores en términos de calidad de audio y precisión de alineación.

Diffsound [YYW+23] es un modelo de generación de texto a sonido que introduce la difusión discreta como enfoque principal para crear sonidos de alta fidelidad a partir de entradas de texto. El objetivo de este estudio es mejorar la calidad y coherencia de la generación de audio condicionada por texto, abordando los problemas asociados con técnicas autorregresivas y optimizando la relación entre el texto de entrada y el audio generado. La innovación principal de este modelo consiste en su decodificador de token no autorregresivo, basado en un modelo de difusión discreta, lo cual le permite predecir todos los tokens del espectrograma de mel en un solo paso, refinándolos iterativamente para obtener resultados de mayor calidad. A diferencia de los decodificadores autorregresivos tradicionales, Diffsound no se limita a predecir tokens secuencialmente, lo que le permite capturar mejor la complejidad de la señal de audio. El modelo está entrenado en conjuntos de datos como Audioset y AudioCaps. Aunque el modelo logra un buen rendimiento en la generación de audio al reducir significativamente los errores acumulados gracias al enfoque no autorregresivo, presenta ciertas limitaciones. El entrenamiento requiere conjuntos de datos extensos y bien etiquetados.

[Bet23] Tortoise es un sistema de síntesis de voz de última generación que destaca por su enfoque en el escalado y la combinación de múltiples modelos para lograr una alta calidad de audio. En lugar de depender de un único modelo complejo, TorToise utiliza cuatro modelos especializados que trabajan en conjunto: Un modelo autorregresivo (AR) que predice una secuencia de tokens de voz discretos a partir del texto de entrada; Un modelo probabilístico de difusión con eliminación de ruido (DDPM) que transforma el espectrograma MEL, condicionado por los tokens de voz del modelo AR; Un codificador de voz discreto/codificador automático variacional (VAE) entrenado en un gran conjunto de datos de audio sin procesar para crear una representación latente comprimida del

audio y un modelo de transformador contrastivo pre-entrenado en lenguaje-voz (CLVP) que funciona como un discriminador cualitativo para re-clasificar las salidas del modelo AR y seleccionar las mejores. Al combinar estos modelos, TorToise aprovecha las ventajas de cada enfoque, logrando una síntesis de voz de alta fidelidad con un control preciso sobre la prosodia y el estilo. Los datasets utilizados para el entrenamiento del modelo fueron LibriTTS y HiFiTTS. El uso de CLVP para la re-clasificación de las salidas del modelo AR permite seleccionar las mejores muestras y mejorar aún más la calidad del audio. Además, el enfoque de escalado utilizado en TorToise, que implica entrenar cada modelo en grandes conjuntos de datos y luego combinarlos, permite una mayor eficiencia y un mejor rendimiento en comparación con los modelos tradicionales de TTS. Aunque TorToise logra una alta calidad de audio, su rendimiento depende en gran medida de la calidad y la cantidad de datos utilizados para entrenar cada modelo. La disponibilidad de conjuntos de datos de voz de alta calidad y de gran tamaño puede ser un factor limitante, especialmente para idiomas con pocos recursos.

Make-An-Audio [HHY+23] es un modelo de difusión mejorado por indicaciones para la generación de texto a audio que aborda los desafíos de la escasez de datos y la complejidad del modelado de audio largo y continuo. Utiliza modelos de difusión mejorados con el poder del pre-entrenamiento contrastivo de lenguaje-audio (CLAP), lo que permite al modelo aprender representaciones robustas de texto y audio sin requerir grandes cantidades de datos anotados. Genera espectrogramas intermedios en lugar de formas de onda directamente, lo que simplifica el proceso de síntesis acústica e implementa un sistema de prompts mejorados, que refina la conexión entre las entradas textuales y las características del audio generado, asegurando una mayor alineación semántica. El modelo fue entrenado con una combinación de diversos datasets como AudioSet,BBC sound effects, Audiostock, AudioCaps-train, ESC-50. Las evaluaciones, tanto subjetivas como objetivas, muestran la capacidad de Make-An-Audio para generar audio fiel a las descripciones de texto. Además, se generaliza bien a múltiples modalidades de entrada (audio, imagen y video), lo que permite a los usuarios crear contenido de audio más diverso. Este modelo presenta una complejidad computacional elevada debido a la arquitectura de difusión y los procesos de preentrenamiento.

3.2.2. Modelos Basados en Modelado de Lenguaje

Google propuso un nuevo marco para la síntesis de audio que utiliza un enfoque de modelado del lenguaje para producir audio realista, incluyendo voz y música de piano, a partir de únicamente señales de audio [BMV+22]. AudioLM se entrena con grandes corpus de formas de onda de audio sin procesar, aprendiendo a generar continuaciones naturales y coherentes a partir de indicaciones cortas. A diferencia de otros modelos que requieren transcripciones de texto o representaciones simbólicas de la música, AudioLM opera directamente sobre la señal de audio. Para lograr esto, utiliza una combinación jerárquica de tokens semánticos y acústicos. Los tokens semánticos capturan la estructura global y las dependencias locales del audio, mientras que los tokens acústicos preservan

los detalles finos de la forma de onda. Al ser condicionado con una breve muestra de audio, AudioLM puede generar continuaciones coherentes que mantienen la identidad del hablante, la prosodia y las condiciones de grabación en el caso del habla, y la melodía, armonía y ritmo en el caso de la música de piano. Los datasest que se utilizaron para el entrenamiento del modelo son Unlab-60 k de Libri-Light. Consta de 60 k horas de habla en inglés. En cuanto a los resultados, AudioLM fue evaluado cualitativamente por las métricas Word Error Rate (WER) y Character Error Rate (CER) y presenta un bajo WER y CER, lo que indica una alta precisión en la transcripción y una efectiva captura del contenido semántico del audio.

AUDIOGEN [KSP+22] tiene la capacidad de generar audio realista y de alta calidad, incluyendo música y efectos de sonido, a partir de descripciones textuales. Utiliza el códec neuronal EnCodec para aprender tokens de audio discretos a partir de la señal sin procesar. Estos tokens proporcionan un ”vocabulario”fijo para representar la información musical, lo cual permite entrenar modelos lingüísticos autorregresivos para generar nuevos tokens, sonidos y música al convertir los tokens de nuevo al espacio de audio con el decodificador de EnCodec. El modelo también utiliza una técnica de aumento que mezcla diferentes muestras de audio, lo que le permite aprender a separar múltiples fuentes de sonido. El modelo está entrenado en grandes conjuntos de datos como AudioSet, BBC sound effects, AudioCaps, y Clotho v2, que contienen ejemplos diversos de audio y descripciones textuales. Soporta una amplia gama de descripciones textuales y genera audio con alta fidelidad y coherencia semántica. Sin embargo, presenta limitaciones en la diversidad cultural representada en los datasets y puede tener dificultades para capturar aspectos acústicos poco comunes o muy complejos.

3.3. Modelos que Combinan Capacidades de Texto a Voz y de Texto a Audio

3.3.1. Modelos que Combinan el Control de los Estilos de la Voz con Contexto Ambiental

Los autores de [LYNC24] proponen VOICELDM, que es un modelo de síntesis de texto a voz con la capacidad de generar audio que refleja con precisión dos indicaciones de texto distintas: un mensaje de descripción y un mensaje de contenido. El mensaje de descripción proporciona información sobre el contexto ambiental general del audio (por ejemplo, ün hombre hablando en una catedral”), mientras que el mensaje de contenido especifica el contenido lingüístico del habla. El texto de entrada se procesa junto con información contextual, como ruidos de fondo o efectos ambientales, para generar audio que refleje no solo el contenido verbal, sino también el entorno deseado. Basado en modelos de difusión latente y aprovechando el preentrenamiento de audio y lenguaje contrastivo (CLAP) y Whisper, VoiceLDM se entrena con grandes cantidades de audio del mundo real sin necesidad de anotaciones o transcripciones manuales. Los datasets utilizandos para el entrenamiento del modelo son AudioSet, el subconjunto en inglés de CommonVoice 13.0

corpus, VoxCeleb y DEMAND. Una de las principales ventajas que ofrece este modelo es la posibilidad de generar audios personalizados mediante la integración del contexto ambiental, con la capacidad de generar una amplia gama de sonidos, incluyendo discursos con efectos de sonido, voces cantadas y susurros, lo que lo convierte en una herramienta versátil para diversas aplicaciones. Una de las principales limitaciones que presenta el modelo es que el entrenamiento requiere grandes cantidades de datos de audio del mundo real y un proceso computacionalmente intensivo. Esto puede limitar la accesibilidad y la capacidad de adaptar el modelo a nuevos idiomas.

Los autores de [LK24] proponen PARLERTTS, un modelo para generar texto a voz que permite controlar diferentes atributos del habla como género, acento, velocidad del habla, tono y condiciones de grabación, mediante el uso de descripciones en lenguaje natural, lo que permite una guía intuitiva del proceso de síntesis de voz. A diferencia de otros métodos, este no se basa en grabaciones de referencia, el modelo depende completamente de la descripción del texto para las condiciones de las características de la voz. Se basan en el etiquetado automático para escalar a grandes conjuntos de datos y utiliza modelos de lenguaje del habla para sintetizar el habla en diferentes estilos y condiciones de grabación controlados mediante lenguaje natural intuitivo. El modelo usa la librería de generación de audio de propósito general AudioCraft, adaptándola para TTS, y emplea Descript Audio Codec (DAC) para proporcionar representaciones de características discretas. Utiliza dos corpus de habla inglesa derivados del proyecto de audiolibros LibriVox: la parte en inglés de Multilingual LibriSpeech (MLS) (45 000 horas) y LibriTTS-R (585 horas). Ambos conjuntos de datos proporcionan transcripciones y una etiqueta para el género generada mediante un modelo predictivo. El modelo fue evaluado tanto con métricas cuantitativas como cualitativas, dando como resultado valores significativamente más altos y cercanos al audio original en las métricas PESQ (3,84 ±0,10), STOI (0,996 ±0,001) y SI-SDR (26,53 ±1,16) y obtuvo una MOS de 3,92 ±0,07.

La Tabla 3.1 presenta un análisis comparativo de algunos de los modelos de generación de texto a audio (TTA) y texto a voz (TTS) más relevantes. Estos modelos emplean diversas arquitecturas, como enfoques autorregresivos, no autorregresivos, basados en la difusión y en redes neuronales, para generar habla, sonidos y música a partir de entradas de texto y audio. Muchos de estos modelos incorporan vocodificadores neuronales, siendo HiFi-GAN el más común. La calidad del audio generado varía según el modelo, lo que se refleja en sus puntuaciones MOS. Mientras que algunos modelos, como AudioLDM y Tortoise, están disponibles como código abierto, el uso de otros está restringido. Esta comparación resalta los avances en la generación de audio mediante inteligencia artificial y muestra las diferencias entre las distintas arquitecturas.

La Tabla 3.2 presenta una comparación de los conjuntos de datos utilizados por cada modelo, algunos son más especializados como LibriTTS y VCTK y otros son de gran tamaño como AudioSet y BBC Sound Effects.

Tabla 3.1: Comparación del estado del arte de los modelos de generación de audio.

Tipo	Modelo	Arquitectura	Entrada	Salida	Vocoder	Puntuación MOS	Código Abierto
TTS	ParlerTTS [LK24]	AudioCraft + Cross-attention + DAC + Decoder-only Transformer	Texto	Voz	-	3,92 ± 0,07	S
	XTTS [CDG+24]	Codificador-Decodificador + VQ-VAE	Texto, Audio	Voz	HiFi-GAN	4,007 ± 0,25	S
	YourTTS [CWS+22]	Codificador-Decodificador + SLC	Texto, Audio	Voz	HiFi-GAN	4,24±0,04 (VCTK)	S
	PromptTTS 2 [LGS+]	Difusión + Codificador de Voz de Referencia + Codificador de Texto + LLM	Texto, Audio	Voz	HiFi-GAN	3,88 ± 0,08	N
	Reinforce-aligner [CLL21]	Generador Conv + MRF + Codificador de Fonemas + Refuerzo + Decodificador	Texto	Voz	HiFi-GAN	4,07 ± 0,04	S
TTA Autoregresivo	AUDIOGEN [KSP+22]	Autoencoder + 1D conv	Texto	Voz, Sonidos, Música		-	N
	SPEAR-TTS [KVB+23]	Transformer	Texto	Voz	HiFi-GAN	4,96±0,02 (LJSpeech)	S
	Tortoise [Bet23]	Codificador-Decodificador + DDPM + CLVP	Texto, Audio	Voz	Univnet	-	S
	CHAR2WAV [SMK+17]	Codificador-Decodificador + SampleRNN	Texto	Voz	Vocodificador neuronal	-	N
	Tacotron [WSRS+17]	RNN + Codificador-Decodificador + Atención	Texto	Espectrograma	Griffin-Lim	3,82±0,085	S
	Deep Voice 2 [GAD+17b]	Basado en CNN	Texto	Voz	WaveNet	2,74 ± 0,35 (VCTK)	N
	Deep Voice 3 [PPG+18]	CNN completo + atención + Seq2Seq	Texto, Audio	Voz	Griffin-Lim, WORLD, WaveNet	3,78 ± 0,30	N
TTA No-autoregresivo	Diffsound [YYW+23]	Codificador de Texto + VQ-VAE + Difusión + Codificador de Voz	Texto	Sonidos	MelGAN	3,56	N
	FastSpeech [RRT+19]	Red de alimentación directa + Transformer	Texto	Voz	WaveGlow	3,84 ± 0,08	S
	Fastpitch [Łań21]	Red de alimentación directa + Transformer	Texto	Voz	WaveGlow	4,080 ± 0,133	S
TTA	AudioLDM [LCY+23]	LDM + CLAP	Texto	Voz, Sonidos	HiFi-GAN	-	N
	Auffusion [XDGL24]	Difusión + LLMs	Texto	Sonidos	HiFi-GAN	-	N
	AudioLM [BMV+22]	Tokens semánticos + acústicos + Transformer	Audio	Voz, Sonidos, Música	SoundStream	-	N

Continúa en la siguiente página

Tipo	Modelo	Arquitectura	Entrada	Salida	Vocoder	Puntuación MOS	Código Abierto
	Make-An-Audio [HHY+23]	Difusión + CLAP	Texto, Audio, Imagen, Video	Sonidos	HiFi-GAN	MOS-Q: 72,5, MOS-F: 78,6	S
TTS/TTA	VOICELDM [LYNC24]	Codificador de Contenido + Durador Diferenciable + CLAP + U-Net + VAE	Texto, Audio	Voz, Sonidos	HiFi-GAN	3,96 (CommonVoice 13.0)	S

Tabla 3.2: Conjuntos de datos utilizados por los modelos de generación de audio.

Modelo	**Conjunto de Datos**
AUDIOGEN [KSP+22]	AudioSet, BBC, AudioCaps, Clotho v2
Diffsound [YYW+23]	AudioSet, AudioCaps
AudioLDM [LCY+23]	AudioSet, AudioCaps, Freesound, BBC Sound Effect library
Auffusion [XDGL24]	AudioCaps, WavCaps, MACS, Clotho, ESC50, UrbanSound, Music Instruments, GTZAN
SPEAR-TTS [KVB+23]	LibriLight, LibriTTS, LJSpeech
Tortoise [Bet23]	LibriTTS, HiFiTTS
VOICELDM [LYNC24]	AudioSet, CommonVoice 13.0, VoxCeleb, DEMAND
ParlerTTS [LK24]	Multilingual LibriSpeech (MLS), LibriTTS-R
XTTS [CDG+24]	LibriTTS, LibriLight, Common Voice
YourTTS [CWS+22]	VCTK, TTS-Portuguese Corpus, fr-FR
PromptTTS 2 [LGS+]	AudioSet, BBC sound effects, Audiostock, AudioCaps-train, ESC-50, FSD50K, Free To Use Sounds, Sonniss Game Effects
CHAR2WAV [SMK+17]	VCTK, DIMEX-100
Tacotron [WSRS+17]	Internal North American English
Deep Voice 2 [GAD+17b]	VCTK, AudioBooks
Deep Voice 3 [PPG+18]	VCTK, LibriSpeech
FastSpeech [RRT+19]	LJSpeech
Fastpitch [Łań21]	LJSpeech
Reinforce-aligner [CLL21]	Korean Language (NIKL) corpus, LJSpeech
AudioLM [BMV+22]	Unlab-60k
Make-An-Audio [HHY+23]	AudioSet, BBC sound effects, Audiostock, AudioCaps-train, ESC-50, FSD50K, Free To Use Sounds, Sonniss Game Effects, We Sound Effects, MACS, Epidemic Sound, Urban Sound 8K, WavText5Ks, LibriSpeech, and Medley-solos-DB

3.4. Síntesis del Capítulo

El objetivo de este capítulo ha sido presentar una revisión de trabajos relacionados presentes en la literatura actual, consolidando el estado del arte en la generación de audio sintético. Los modelos de TTS, como SPEAR-TTS, XTTS y FastPitch, han logrado avances significativos en la naturalidad del habla, la capacidad para generar voz en múltiples idiomas y el control prosódico, aunque persisten limitaciones en la escalabilidad para idiomas con pocos recursos y la clonación precisa de voces diversas. Por su parte, los modelos de TTA, como Auffusion y AudioLM, destacan por su capacidad para generar sonidos variados y contextos acústicos, pero enfrentan desafíos en la alineación texto-audio y la diversidad de datos. Los sistemas combinados TTS/TTA, como VOICELDM y ParlerTTS, muestran un potencial innovador al integrar control de estilos vocales y ambientes, sin embargo, requieren grandes volúmenes de datos y recursos computacionales. Las tablas comparativas de modelos y conjuntos de datos muestran la diversidad de arquitecturas y la dependencia de datasets extensos.

Capítulo 4

Propuesta y Metodología

En este capítulo se presenta la propuesta para la generación de audio sintético multilingüe y multiacento, junto con la metodología empleada para su desarrollo. La Sección 4.1 describe la metodología SEMMA, un marco ampliamente utilizado en ciencia de datos, que ha sido implementado en este trabajo para seguir de forma estructurada las etapas de muestreo, exploración, modificación, modelización y evaluación de los modelos. Sección 4.2, se detallan las generalidades de la propuesta, que integra los modelos VoiceLDM y XTTS para generar audios personalizados con control sobre contenido lingüístico, contexto ambiental y características vocales, explicando paso a paso el enfoque propuesto, desde el preprocesamiento de texto hasta la síntesis multilingüe.

4.1. Base Metodológica Seguida

El desarrollo teórico de este trabajo se fundamenta en la metodología SEMMA, un marco ampliamente utilizado en ciencia de datos que estructura el proceso en cinco etapas: Sample (muestreo), Explore (exploración), Modify (modificación), Model (modelado) y Assess (evaluación). En este contexto, la metodología SEMMA se adapta al enfoque de generación de audio sintético multilingüe y multiacento, proporcionando una guía para la preparación de datos, el desarrollo de modelos y la evaluación de resultados. A continuación, se describe cómo se aplican estas etapas en la propuesta de este trabajo:

1. **Muestreo:** Selección de datasets representativos como Common Voice 18.0, LibriTTS y CML-TTS, que garantizan alta calidad, diversidad lingüística y variabilidad de hablantes para entrenar y evaluar los modelos.

2. **Explorar:** Análisis inicial de las características de los datos de audio, como la calidad de las grabaciones, la diversidad de hablantes y las propiedades lingüísticas, para comprender el alcance de los datos disponibles.

3. **Modificar:** Preprocesamiento de los datos, incluyendo, la normalización, tokenización y la eliminación de caracteres especiales, para asegurar que los datos de entrada sean consistentes y adecuados para el entrenamiento y generación de audio.

4. **Modelizar:** Desarrollo y ajuste (fine-tuning) de los modelos VoiceLDM y XTTS, junto con la implementación de un discriminador basado en Wav2Vec 2.0 para validar que el género y la edad proporcionados en la entrada sean coherentes con el audio generado.

5. **Evaluar:** Evaluación de los resultados mediante métricas cuantitativas como Fréchet Audio Distance (FAD), Fréchet Distance (FD), Kullback-Leibler Divergence (KL), CLAP Score y Word Error Rate (WER), comparando el rendimiento con modelos del estado del arte.

4.2. Generalidades de la Propuesta

La propuesta integra los modelos de aprendizaje profundo VoiceLDM [LYNC24] y XTTS [CDG+24] para generar audios sintéticos personalizados con control preciso sobre el contenido lingüístico y las características vocales, como edad, género, idioma y acento. Esta combinación permite producir audios de alta calidad con mayor flexibilidad que los modelos existentes en el estado del arte, superando sus limitaciones en la personalización y diversidad vocal. A continuación, se describe el proceso general seguido por el sistema, resumido en la figura 4.1.

Primero, el sistema recibe como entrada una descripción textual que incluye dos componentes:

- El contexto ambiental y los atributos de la voz deseada (por ejemplo, “una mujer en sus veintes”).

- El contenido lingüístico a expresar (por ejemplo, “Hola, hoy es un día soleado”).

Estas entradas de texto son preprocesadas para eliminar caracteres inválidos y posteriormente se utilizan en un proceso en dos etapas:

1. **VoiceLDM** genera un audio base coherente con las entradas proporcionadas y refleja tanto el contenido lingüístico como las características ambientales especificadas.

2. **XTTS** transforma ese audio base generado por VoiceLDM para adaptarlo a múltiples idiomas, preservando las características vocales originales de la voz.

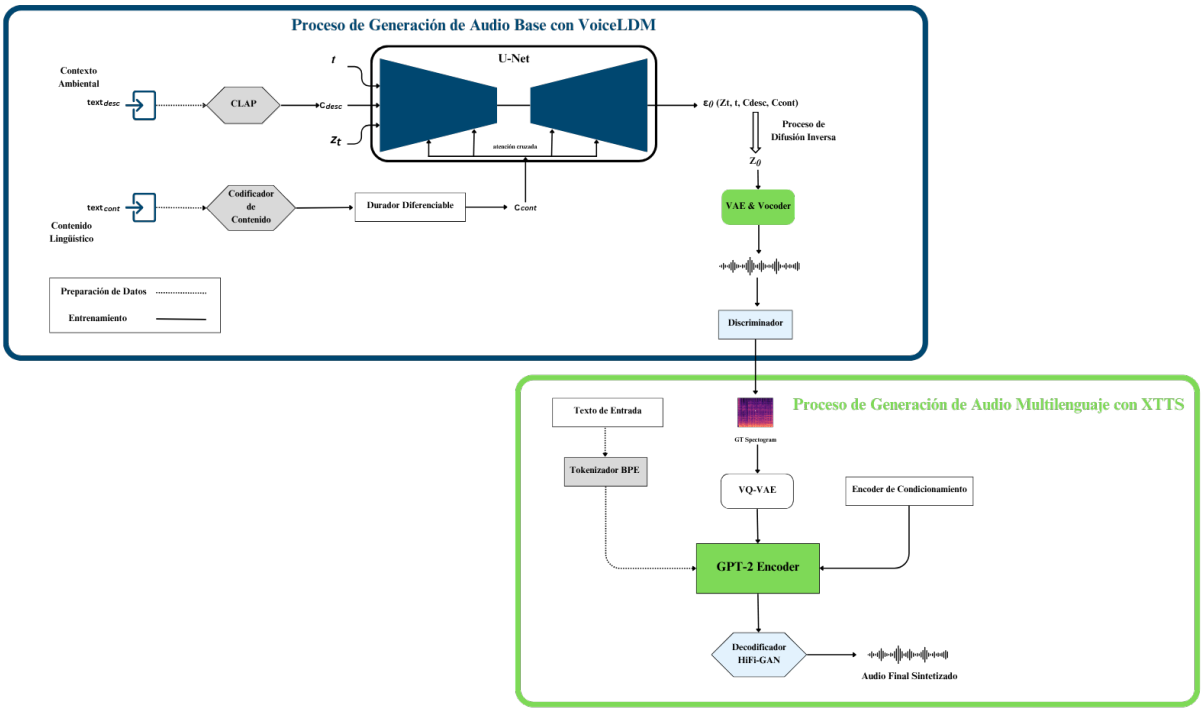


Figura 4.1: Enfoque propuesto para la generación de audio sintético.

4.3. Fine-tuning del Modelo VoiceLDM con Audio Limpio

Para mejorar la calidad del audio generado, se realizó un proceso de fine-tuning al modelo VoiceLDM-M utilizando audio limpio, utilizando datos de alta calidad de los conjuntos de datos Common Voice 18.0 [ABD+19], LibriTTS [KZK+23] y CML TTS [OCJ+23], elegidos para garantizar la alta calidad del audio, la diversidad lingüística y la variabilidad de los hablantes. Los audios de LibriTTS-R y CML TTS proceden de audiolibros de LibriVox, lo que proporciona grabaciones de alta calidad y una amplia cobertura fonética. Por su parte, Common Voice 18.0 es un conjunto de datos de Mozilla que procede de una comunidad mundial de hablantes, ofrece una amplia gama de diversidad de hablantes y variación de acentos. Estos datasets ofrecen alta calidad, diversidad lingüística y variabilidad de hablantes, como se detalla en la tabla 4.1 y en la tabla 4.2 se resume la duración y el número de hablantes por idioma.

Durante la preparación de los conjuntos de datos para el reentrenamiento del modelo, se llevó a cabo un preprocesamiento, el cual incluyó:

- La identificación y eliminación de caracteres inválidos.
- Conversión de números a palabras.
- Limpieza del texto y eliminación de caracteres especiales.
- Tokenización para dividir el texto en unidades compatibles con el modelo.

Tabla 4.1: Conjuntos de Datos.

Conjunto de Datos	**Duración (horas)**	**Número de hablantes**
Common Voice 18.0 (Inglés)	2,640	93,279
Common Voice 18.0 (Español)	2,230	26,107
Common Voice 18.0 (Portugués)	214	3,516
LibriTTS-R (train-960)	585	2,456
CML-TTS (Español)	453.8	115
CML-TTS (Portugués)	69.35	48
TOTAL	6,192.15	125,521

Tabla 4.2: Conjuntos de Datos por Idioma.

Idioma	**Duración (horas)**	**Número de hablantes**
Inglés	3,225	95,735
Español	2,683.8	26,115
Portugués	283.35	3,564
TOTAL	6,192.15	125,521

Este preprocesamiento asegura que el modelo sea entrenado con datos de texto consistentes y limpios, lo que permite que el modelo aprenda a generar audio natural y de mayor calidad, alineado con los datos de entrenamiento optimizados.

Para el ajuste fino, se definieron hiperparámetros clave que controlan el entrenamiento, como el batch size, el learning rate y el número de pasos, los cuales se presentan en la Tabla 3. Estos parámetros fueron ajustados para maximizar la eficiencia computacional y la calidad del audio generado.

Tabla 4.3: Hiperparámetros de entrenamiento del modelo.

Hiperparámetro	**Valor**
`max_train_steps`	1,000,000
`checkpointing_steps`	20,000
`train_batch_size`	4
`gradient_accumulation_steps`	1
`total_batch_size`	16
`dataloader_num_workers`	8
`learning_rate`	2×10^{-5}

4.4. Generación de Audio Sintético con VoiceLDM

El proceso comienza con dos entradas principales:

- **textdesc:** descripción ambiental que define el contexto en el que se generará el audio y los atributos de la voz deseados.

- **textcont:** contenido lingüístico a ser convertido en audio.

Antes de ser procesadas por el modelo, ambas entradas son sometidas a una etapa de preprocesamiento textual, con el objetivo de garantizar su compatibilidad con el modelo y mejorar la calidad fonética del audio generado. En la Figura 4.2 se ilustra todo el proceso que sigue, desde las entradas de texto hasta la generación del audio base en inglés, español o portugués.

Este preprocesamiento incluye las siguientes operaciones:

1. **Normalización de caracteres**: Se reemplazan acentos y caracteres por sus equivalentes básicos (por ejemplo, "ñ" por "n", "ç" por "c", "ü" por "u").
2. **Conversión de números a palabras**: se utilizan reglas específicas por idioma (*en*, *es*, *pt*) mediante el uso de la librería `num2words` para transformar cifras numéricas en su representación textual.
3. **Sustitución de símbolos especiales**: Se realiza el remplazo de símbolos especiales por idioma. Por ejemplo, caracteres como `$`, `%`, `#`, `@`, entre otros, se reemplazan por expresiones textuales correspondientes (Por ejemplo, en español y portugués el símbolo "@" se transforma en "arroba", mientras que en inglés se reemplaza por "at").
4. **Procesamiento de direcciones de correo electrónico**: Se identifican expresiones con formato de email y se transforman en texto explícito, reemplazando "@" por "arroba" y "." por "punto".
5. **Eliminación de caracteres no válidos**: Se filtran y eliminan todos aquellos símbolos no reconocidos por el modelo.
6. **Eliminación de espacios redundantes**: Se normaliza el texto final para asegurar una estructura coherente y libre de espacios innecesarios.

Una vez completado el preprocesamiento del texto, las entradas se procesan en varias etapas para garantizar que el audio generado refleje fielmente tanto el contenido lingüístico como el contexto ambiental deseado, como se explica a continuación:

4.4.1. Codificación del Texto

- **textdesc** se codifica mediante el modelo CLAP [WCZ+23], el cual proyecta texto y audio en un espacio latente, lo que permite transformar `textdesc` en un vector latente $\mathbf{c}_{\text{desc}} \in \mathbb{R}^{512}$.
- **textcont** es procesado por un codificador que genera una secuencia oculta $\mathbf{H}_{\text{cont}} \in \mathbb{R}^{L \times D}$, donde L es la longitud de la secuencia y D es la dimensión de cada representación.

Estas representaciones latentes contienen información relevante extraída del texto de entrada, como la semántica, las relaciones contextuales y las dependencias presentes en el texto, y son utilizadas en etapas posteriores para guiar la síntesis del audio.

4.4.2. Sincronización del Contenido con la Duración del Audio

Para sincronizar la duración del contenido lingüístico con la del audio generado, se utiliza un módulo de duración diferenciable, con el fin de garantizar que el texto codificado ($\mathbf{H}_{\text{cont}}$) se ajuste correctamente con la duración esperada del audio final.

Para adaptarse a las variaciones naturales del habla, como pausas y ritmos, el módulo transforma $\mathbf{H}_{\text{cont}} \in \mathbb{R}^{L \times D}$ en una secuencia temporal ($\mathbf{c}_{\text{cont}} \in \mathbb{R}^{N \times D}$), donde $N \geq L$, permitiendo ajustar la longitud del audio según las necesidades temporales del habla.

Durante este proceso, el durador realiza un *upsampling* de $\mathbf{H}_{\text{cont}}$, repitiendo ciertos vectores en $\mathbf{c}_{\text{cont}}$ para extender la duración de algunas unidades lingüísticas según sea necesario, y asegurar que las pausas, ritmos y duraciones sean consistentes con el habla natural.

4.4.3. Síntesis del Audio por Difusión Latente

La red generativa principal del sistema es una arquitectura *U-Net* que recibe cuatro entradas:

- $\mathbf{c}_{\text{desc}}$: vector que representa el contexto ambiental y las características de la voz.
- $\mathbf{c}_{\text{cont}}$: contenido lingüístico alineado temporalmente.
- $\mathbf{z}_t$: representación latente ruidosa.
- t: incrustación temporal que indica la etapa actual del proceso de difusión.

El objetivo de la *U-Net* es predecir la puntuación de difusión ($\boldsymbol{\varepsilon}_\theta$) necesaria para guiar el proceso de difusión inversa, el cual comienza desde el ruido aleatorio y va reconstruyendo el audio, integrando el contexto y el contenido. Este proceso es iterativo y continúa hasta que se alcanza una representación final del audio, alineado con el contenido lingüístico y el contexto ambiental especificados.

La *U-Net* predice la puntuación de difusión $\boldsymbol{\varepsilon}_\theta$, necesaria para guiar el proceso de difusión inversa. Este proceso comienza con ruido aleatorio y de forma iterativa, va eliminando el ruido para generar una representación latente $\mathbf{z}_0$. Durante cada paso, se integran $\mathbf{c}_{\text{desc}}$ y $\mathbf{c}_{\text{cont}}$ para asegurar que la representación latente resultante sea coherente con el contenido lingüístico y el contexto ambiental.

4.4.4. Decodificación y Vocoder

El resultado del proceso de difusión inversa ($\mathbf{z}_0$) es una representación latente del audio, la cual se decodifica utilizando un codificador automático variacional (*VAE*) preentrenado para obtener un espectrograma de Mel.

Este espectrograma se convierte en audio mediante el vocoder *HiFi-GAN*, garantizando alta fidelidad y naturalidad en el resultado final.

4.4.5. Validación de Edad y Género del Audio Base con Discriminador

Para garantizar que el audio base generado por *VoiceLDM* cumpla con las características deseadas del hablante especificadas en la entrada `textdesc` (como el rango de edad y el género, por ejemplo, "una mujer en sus veintes"), se implementa un mecanismo de validación posterior a la generación del audio base. Este mecanismo utiliza un discriminador basado en *Wav2Vec* [BWW+23], un modelo preentrenado que ha sido ajustado específicamente para estimar edad y género a partir de señales de audio.

El discriminador recibe el audio base y predice:

- **Edad**: valor continuo en el rango $[0, 1]$, mapeado a 0–100 años.
- **Género**: probabilidades para *niño*, *mujer*, *hombre*.

La predicción se compara con las características esperadas definidas en la entrada `textdesc`. Si el resultado no cumple con los criterios establecidos, el sistema descarta el audio y reinicia el proceso de generación con el mismo contenido y descripción, hasta obtener un resultado válido.

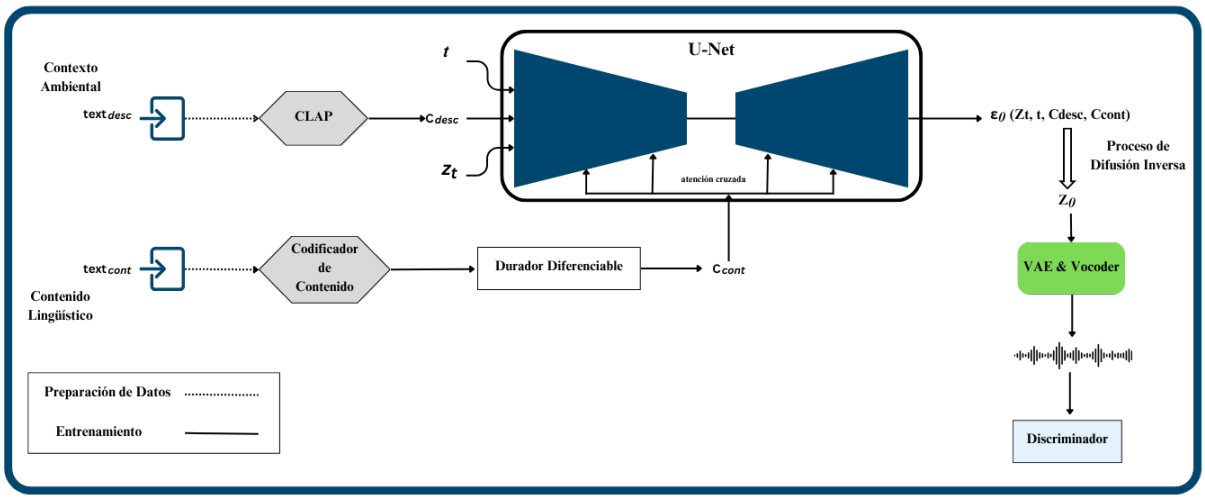


Figura 4.2: Proceso para la generación de audio sintético con VoiceLDM.

Una vez generado el audio sintético base con el modelo *VoiceLDM* y realizada la comprobación de coherencia con la edad y el género proporcionados en la entrada,

la siguiente fase se centra en transformar este audio para darle mayor versatilidad. Aunque el audio base ya refleja tanto el contenido lingüístico como el contexto ambiental especificados, su alcance está limitado a los idiomas inglés, español y portugués.

Para afrontar estas limitaciones, se utiliza el modelo *XTTS*, especializado en la síntesis de voz multilingüe, para generar nuevos audios en distintos idiomas, conservando las características vocales del hablante. A continuación, se describe el proceso de adaptación llevado a cabo con *XTTS*.

4.5. Generación de Audio Sintético en Múltiples Lenguajes con XTTS

Una vez generado el audio base con *VoiceLDM*, se emplea el modelo *XTTS* para transformar dicho audio y adaptarlo a distintos idiomas, preservando las características vocales del hablante. La Figura 4.3 ilustra el flujo de este proceso.

Para lograr este objetivo, se implementa un flujo de trabajo que consta de las siguientes fases:

4.5.1. Procesamiento Inicial del Audio de Referencia

Esta etapa se enfoca en extraer las características acústicas y de voz del audio de referencia generado por *VoiceLDM*, las cuales serán clonadas en la salida.

- **Codificación con *Autocodificador Variacional Cuantificado Vectorial* (VQ-VAE)**: El audio de referencia de entrada se convierte primero en un mel-espectrograma y luego se procesa mediante un VQ-VAE, Este componente comprime la señal acústica en una secuencia de códigos de audio discretos. Para mejorar la expresividad del modelo, XTTS filtra el libro de códigos del VQ-VAE, conservando únicamente los 1024 códigos más frecuentes de 8192 códigos posibles.

- **Extracción de Incrustaciones de Condicionamiento**: Paralelamente, se utiliza un Encoder de Condicionamiento que recibe el mel-espectrograma del audio de referencia, el cual genera 32 incrustaciones (embeddings) de 1024 dimensiones para cada muestra de audio. Su propósito es producir incrustaciones que condicionan el GPT-2 Encoder.

- **Extracción de Incrustaciones del Hablante (H/ASP Model)**: La incrustación del hablante se obtiene del audio de referencia utilizando el modelo H/ASP (Hidden/Auxiliary Speaker Prior). Esta incrustación captura las características específicas de la voz, como el timbre, el estilo del habla o la personalidad vocal, y es fundamental para la clonación de la voz del hablante de referencia en el audio final.

4.5.2. Procesamiento de Texto y Generación de Vectores Latentes (Encoder GPT-2)

Esta es la fase central donde el texto de entrada se transforma en representaciones que serán decodificadas en voz. El componente principal es un encoder basado en GPT-2 que recibe como entrada principal tokens de texto obtenidos mediante un tokenizador personalizado de codificación por pares de bytes (BPE) de 6681 tokens. El encoder GPT-2 es condicionado por las incrustaciones de audio (las 32 incrustaciones de 1024 dimensiones) generadas por el Conditioning Encoder, como se explicó en la Sección 4.5.1. El encoder GPT-2 predice los códigos de audio VQ-VAE. Sin embargo, para la decodificación subsiguiente, se utilizan los vectores latentes del espacio latente del encoder GPT-2, en lugar de los códigos VQ-VAE directos. Esto se hace para evitar problemas de pronunciación debido a la alta tasa de compresión del VQ-VAE.

4.5.3. Síntesis de Audio Final (Decodificador HiFi-GAN)

La última etapa se encarga de convertir los vectores latentes en una señal de audio de alta fidelidad. El componente principal es un decodificador basado en el vocoder HiFi-GAN, el cual recibe como entrada los vectores latentes del encoder GPT-2 y la incrustación del hablante del modelo H/ASP (Sección 4.5.2) y sintetiza el audio final, el cual imita la voz del hablante de referencia y pronuncia el texto de entrada en el idioma deseado.

4.6. Síntesis del Capítulo

Este capítulo ha consolidado la propuesta metodológica para la generación de audio sintético multilingüe y multiacento, destacando la utilidad de la metodología SEMMA como guía sistemática para estructurar el flujo de trabajo, desde la preparación de los datos hasta la evaluación de los modelos. La integración de VoiceLDM y XTTS permite generar audios personalizados de alta calidad, combinando control preciso sobre el contenido lingüístico, el contexto ambiental y las características de la voz, superando limitaciones de modelos previos en flexibilidad y diversidad. El proceso detallado, desde el preprocesamiento del texto hasta la síntesis multilingüe con validación de edad y género mediante un discriminador basado en Wav2Vec 2.0, muestra coherencia y naturalidad en los resultados. Como resultado de esta investigación, se elaboró y envió un artículo científico basado en los principales aportes técnicos y experimentales del trabajo. La información detallada sobre este envío se encuentra disponible en el Anexo C.

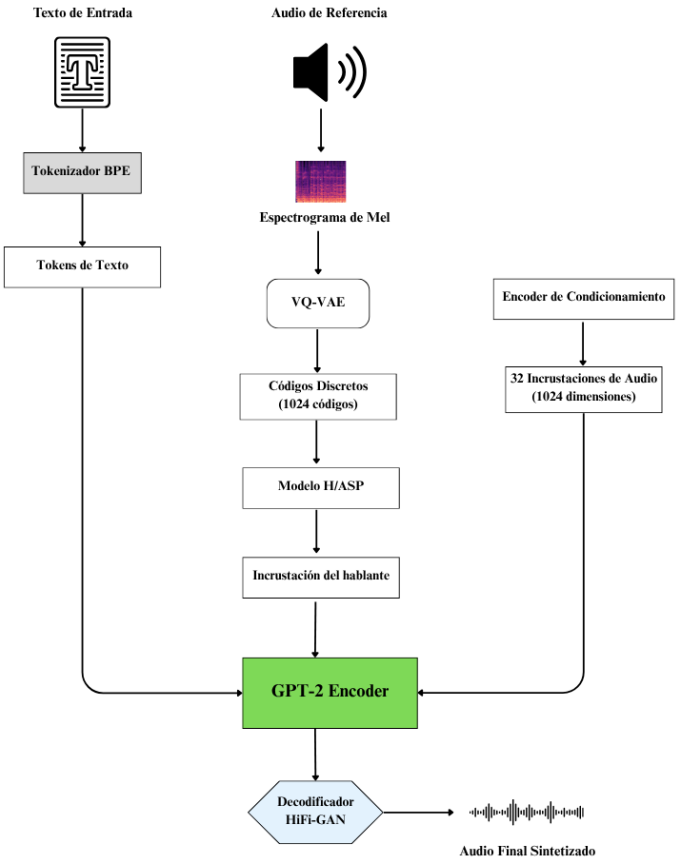


Figura 4.3: Generación de Audio Multilenguaje con XTTS a partir del Audio de Referencia.

Capítulo 5

Experimentos y Resultados

En este capítulo se describe la evaluación experimental del enfoque propuesto para la generación de audio sintético multilingüe y multiacento. En la Sección 5.1, se detalla la configuración del entorno computacional utilizado para los experimentos, asegurando la reproducibilidad de los resultados. En la Sección 5.2, se presentan las métricas cuantitativas empleadas para medir la calidad, inteligibilidad y coherencia contextual del audio generado. En la Sección 5.3, se analizan los resultados del modelo VoiceLDM antes y después del fine-tuning. Finalmente, en la Sección 5.4, se comparan los resultados del enfoque integrado VoiceLDM + XTTS con modelos del estado del arte.

Para evaluar la eficacia del enfoque propuesto para la generación de audio sintético en dos fases, se ha llevado a cabo una serie de experimentos utilizando el modelo VoiceLDM (base y tras fine-tuning) y su integración con XTTS para la generación multilingüe. El objetivo es comprobar la calidad, naturalidad, coherencia contextual y precisión lingüística del audio generado. Se utilizaron varias métricas cuantitativas que permiten evaluar diferentes dimensiones de la calidad del audio, y se comparó el rendimiento de los modelos propuestos con otros enfoques existentes, así como con muestras reales (ground truth).

5.1. Configuración del Entorno

Todos los experimentos fueron realizados en un entorno de computación configurado para soportar modelos de generación de audio sintético de alta demanda computacional, como VoiceLDM y XTTS. Se utilizó un sistema equipado con un procesador de alto rendimiento (equivalente a un Intel Core i5-13600K o AMD Ryzen 7 7700), al menos 16 GB de memoria RAM, una GPU compatible con CUDA (mínimo 8 GB de VRAM) y un disco con 50 GB de espacio libre, operando bajo Ubuntu 20.04 o superior. El entorno de software se basó en Python (versión 3.9 o superior) como lenguaje principal, con soporte para CUDA 11.8 o superior. Las librerías utilizadas incluyeron PyTorch (2.5.1), torchaudio (2.5.1), TTS (0.22.0), accelerate (1.3.0), diffusers (0.32.2), transformers (4.48.1), scipy (1.15.1) y librosa (0.10.2). Esta configuración aseguró la ejecución eficiente de los procesos de entrenamiento, ajuste y generación de audio sintético descritos en este trabajo.

5.2. Métricas de Evaluación

Para evaluar el rendimiento del enfoque propuesto de generación de audio sintético multilingüe y multiacento en dos fases, se emplearon métricas cuantitativas que miden diferentes aspectos del audio generado: calidad acústica, inteligibilidad y coherencia contextual. Las métricas seleccionadas incluyen *Frechet Audio Distance (FAD)*, *Frechet Distance (FD)*, *Kullback-Leibler Divergence (KL)*, *CLAP Score* y *Word Error Rate (WER)*. Estas métricas permiten una evaluación objetiva y comprensiva, comparando los audios generados con el *ground truth* y modelos del estado del arte. A continuación, se detalla cada métrica:

5.2.1. Métricas de Calidad de Audio

- **Fréchet Audio Distance (FAD)**:
  La FAD, también conocida como distancia de Fréchet o distancia de Wasserstein-2, mide la similitud estadística entre las distribuciones de audio generado y audio real. Se calcula extrayendo incrustaciones de audio a partir de un enfoque basado en características acústicas, y luego determinando la distancia de Fréchet entre las distribuciones resultantes [Sum24]. Un valor más bajo indica mayor similitud acústica y, por tanto, mejor calidad del audio generado. La fórmula de la FAD entre dos distribuciones gaussianas $N(\mu_r, \Sigma_r)$ (audio real) y $N(\mu_g, \Sigma_g)$ (audio generado) es:

$$\text{FAD} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}\right), \tag{5.1}$$

  donde μ_r, μ_g son las medias, Σ_r, Σ_g son las matrices de covarianza, y Tr denota la traza.

- **Fréchet Distance (FD)**:
  La FD mide la distancia entre las distribuciones de audio real y generado en un espacio de características, de manera similar a la FAD pero utilizando diferentes enfoques para la extracción de incrustaciones, como incrustaciones extraídas de modelos preentrenados específicos, como PANNs (un modelo de clasificación de audio de última generación) [KCI+20]. Al igual que la FAD, se calcula con la fórmula (5.1) [Sum24], pero las incrustaciones de PANNs capturan patrones acústicos más específicos, como eventos sonoros complejos. Un valor más bajo de FD refleja mayor similitud entre las distribuciones, indicando una mejor calidad del audio generado.

- **Kullback-Leibler Divergence (KL)**:
  La divergencia de Kullback-Leibler mide la diferencia entre dos distribuciones de probabilidad, en este caso, entre las distribuciones de audio generado y real. Se calcula a partir de incrustaciones de audio que representan las características acústicas, evaluando la diferencia entre ambas distribuciones. Un valor más bajo indica mayor similitud entre las distribuciones [Sum24]. La fórmula general para la KL entre dos distribuciones P (real) y Q (generado) es:

$$D_{KL}(P\|Q) = \sum_{x} P(x) \log\left(\frac{P(x)}{Q(x)}\right). \tag{5.2}$$

En este trabajo, las incrustaciones se normalizan con funciones sigmoide o softmax antes de calcular la divergencia, lo que permite evaluar la naturalidad acústica en diferentes niveles de abstracción.

5.2.2. Métricas de Inteligibilidad y Precisión

- **Word Error Rate (WER)**:
  El WER mide la precisión lingüística del audio generado al compararlo con el texto objetivo. Se calcula transcribiendo el audio generado con un modelo de reconocimiento automático del habla (ASR), en este caso Wav2Vec 2.0, y comparando la transcripción con el texto de referencia. El WER se define como:

$$\text{WER} = \frac{S + D + I}{N} \times 100, \tag{5.3}$$

  donde S es el número de sustituciones, D el número de eliminaciones, I el número de inserciones, y N el número total de palabras en el texto de referencia. Un valor más bajo indica una mejor alineación del contenido lingüístico generado con el texto deseado.

5.2.3. Métricas de Coherencia Contextual

- **CLAP Score**:
  El CLAP Score mide la coherencia entre el texto de entrada (por ejemplo, contexto ambiental como "a man in his twenties talking") y el audio generado, utilizando el modelo CLAP. Se calcula mediante la similaridad coseno entre las incrustaciones del texto y el audio. La fórmula de la similaridad coseno entre dos vectores $\mathbf{t}$ (texto) y $\mathbf{a}$ (audio) es:

$$\text{CLAP Score} = \frac{\mathbf{t} \cdot \mathbf{a}}{\|\mathbf{t}\| \, \|\mathbf{a}\|}. \tag{5.4}$$

  Un valor más alto (cercano a 1) indica mayor coherencia entre el contexto especificado y el audio generado. Esta métrica es crucial para evaluar el control contextual del sistema, especialmente en datasets como Common Voice 18.0, donde se especifican contextos ambientales.

5.3. Evaluación del modelo VoiceLDM tras fine-tuning

Al modelo VoiceLDM se le realizó un proceso de fine-tuning con el objetivo de mejorar su capacidad para generar audios de alta calidad, naturales y alineados con el contenido textual. Para ello, se utilizaron subconjuntos de test de los datasets *Common Voice 18.0*

(en inglés, español y portugués), *LibriTTS-R (train-960)* y *CML-TTS* (en español y portugués), seleccionados por su diversidad de hablantes, riqueza fonética y calidad de grabación.

Durante las pruebas, se aplicaron las métricas descritas previamente (FAD, FD, KL, CLAP y WER) para evaluar el rendimiento del modelo antes y después del fine-tuning. La generación de audio se realizó utilizando el contexto ambiental .ª man/woman in his/her teens/twenties/.../nineties talking"para el dataset Common Voice (que incluye metadatos de edad y género) y el contexto genérico çlean speech"para los demás conjuntos.

Los resultados obtenidos tras el fine-tuning se presentan en la Tabla 5.1. El modelo mejoró significativamente en todas las métricas evaluadas:

Tabla 5.1: Evaluación del modelo VoiceLDM antes y después del fine-tuning comparado con el ground truth en los datasets Common Voice 18.0 (inglés, español y portugués), LibriTTS-R (train-960) y CML-TTS (español y portugués). ↑: mayor es mejor; ↓: menor es mejor.

Modelo	**FAD ↓**	**KL ↓**	**FD ↓**	**CLAP Score ↑**	**WER (%) ↓**
VoiceLDM (base)	6,317	0,007	7,723	0,179	47,9
VoiceLDM (fine-tuned)	**5,261**	**0,006**	**6,687**	**0,213**	**31,8**
Ground Truth	—	—	—	0,273	0,0

Los resultados muestran una mejora clara en la similitud estadística con el audio real (reducción de FAD de 6,317 a 5,261 y de FD de 7,723 a 6,687), una menor divergencia entre distribuciones (KL de 0,007 a 0,006), un incremento en la coherencia semántica con el texto (CLAP Score de 0,179 a 0,213), y una mejora considerable en la inteligibilidad del contenido (reducción del WER del 47,9 % al 31,8 %).

En particular, la reducción del WER del 47,9 % al 31,8 % demuestra que el fine-tuning mejora la precisión del contenido lingüístico generado por el modelo. El aumento del CLAP Score indica una mayor fidelidad entre el audio generado y el contexto ambiental definido, acercándose al Ground Truth.

En conjunto, estos resultados indican que el fine-tuning del modelo VoiceLDM sobre datos limpios y representativos permite generar audios más naturales, coherentes y fieles al texto de entrada, con mejoras notables en calidad acústica, inteligibilidad y alineación semántica.

5.4. Evaluación del Enfoque Propuesto

5.4.1. Comparación Cuantitativa con Modelos Existentes

Con el objetivo de evaluar el rendimiento del enfoque propuesto frente a otros modelos del estado del arte, se ha llevado a cabo una comparación cuantitativa utilizando las métricas descritas previamente.

Para garantizar una comparación justa y consistente, las métricas fueron calculadas a partir de audios generados siguiendo las mismas condiciones de prueba, es decir, utilizando el mismo conjunto de datos y criterios para todos los modelos. En concreto, todos los modelos—incluidos **fairseq** y **Parler-TTS**—fueron evaluados utilizando los subconjuntos de prueba de los datasets **Common Voice 18.0** (en inglés, español y portugués), **LibriTTS-R (train-960)** y **CML-TTS** (en español y portugués). Esto asegura que las diferencias observadas en los resultados reflejan únicamente el comportamiento de los modelos, y no variaciones en los datos de entrada.

En la Tabla 5.2 se resume los resultados obtenidos por los modelos **fairseq**, **Parler-TTS**, **VoiceLDM** (modelo base y ajustado), y el enfoque propuesto **VoiceLDM + XTTS** tras el fine-tuning. También se incluyen los valores de referencia correspondientes al **ground truth**. Los valores se han redondeado para facilitar su interpretación.

Tabla 5.2: Comparación del rendimiento entre modelos del estado del arte con métricas cuantitativas en los datasets Common Voice 18.0 (inglés, español y portugués), LibriTTS-R (train-960) y CML-TTS (español y portugués). ↑: mayor es mejor; ↓: menor es mejor.

Modelo	**FAD ↓**	**KL ↓**	**FD ↓**	**CLAP Score ↑**	**WER (%) ↓**
fairseq	6,456	**0,002**	2,055	0,188	21,5
Parler-TTS	**1,406**	0,003	**0,584**	0,218	20,4
VoiceLDM (base)	6,317	0,007	7,723	0,179	47,9
VoiceLDM (fine-tuned)	5,261	0,006	6,687	0,213	31,8
VoiceLDM + XTTS	2,082	0,003	1,406	**0,229**	**18,5**
Ground Truth	—	—	—	0,273	0,0

Los resultados obtenidos evidencian que el enfoque propuesto, compuesto por los modelos **VoiceLDM + XTTS**, logra un rendimiento destacado en términos de coherencia semántica y fidelidad lingüística. En particular, alcanza el **mejor valor de WER** (18,5 %) y el **CLAP Score más alto** (0,229), siendo este último el más cercano al valor del **ground truth** (0,273). Esto indica que el audio generado por el enfoque propuesto refleja con mayor precisión el contenido textual y el contexto deseado.

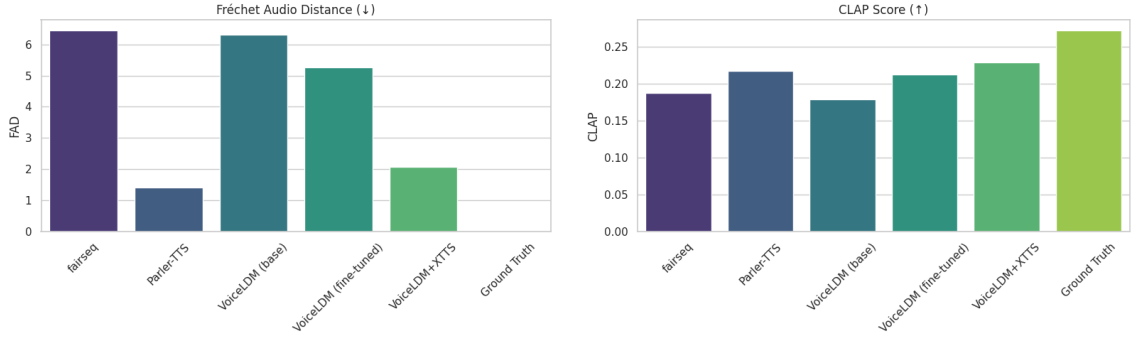
Aunque **Parler-TTS** obtiene los mejores resultados en las métricas de la calidad del audio como **FAD** (1,406) y **FD** (0,584), su desempeño en CLAP Score (0,218) y WER (20,4 %) es inferior al del enfoque propuesto. Esto sugiere que, si bien Parler-TTS genera audio perceptivamente natural, no logra capturar con la misma eficacia el contenido semántico y lingüístico del texto de entrada.

En comparación con **fairseq**, el enfoque propuesto muestra mejoras significativas en todas las métricas clave: reduce tanto FAD como FD, mejora el CLAP Score y disminuye el WER, lo que indica una mayor similitud estadística y contextual con el habla natural.

Finalmente, respecto al modelo **VoiceLDM (base)**, las mejoras tras el *fine-tuning* son evidentes. El FAD disminuye de 6,317 a 2,082, el FD se reduce de 7,723 a 1,406, el WER baja de 47,9 % a 18,5 %.

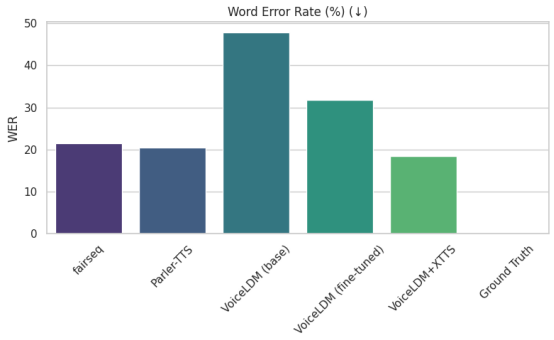
Las mejoras son evidentes, validando la eficacia del fine-tuning y la integración con XTTS para obtener mejores resultados. Esto confirma que la combinación VoiceLDM + XTTS no solo mejora la calidad semántica y lingüística, sino que también representa una solución eficaz frente a otras alternativas actuales en tareas de generación de audio sintético multilingüe.

Para visualizar estas diferencias de manera más clara, en la Figura 5.1 se presentan gráficos de barras que comparan el rendimiento de todos los modelos en tres métricas clave: FAD, CLAP Score y WER. Estas visualizaciones permiten identificar fácilmente los puntos fuertes de cada modelo en función de la métrica evaluada.



((a)) Frechet Audio Distance (FAD, ↓)

((b)) CLAP Score (↑)



((c)) Word Error Rate (WER, ↓)

Figura 5.1: Comparación de modelos en FAD, CLAP Score y WER.

Los gráficos de barras en la Figura 5.1 permiten observar las diferencias entre los modelos en las métricas FAD, CLAP Score y WER.

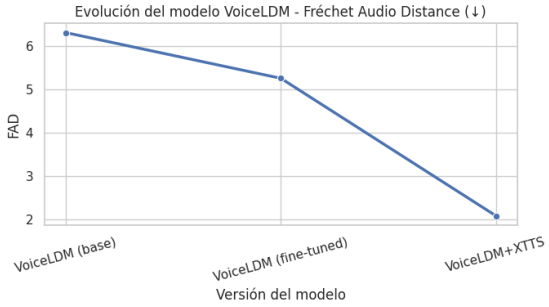
En el Gráfico de FAD, la barra correspondiente a **Parler-TTS** es notablemente la más baja, mostrando el mejor valor (1,406), seguida por **VoiceLDM + XTTS** (2,082). Las barras de **VoiceLDM base** (6,317) y **fairseq** (6,456) son las más altas, indicando un desempeño inferior en calidad acústica.

En el Gráfico de CLAP Score, las barras reflejan un aumento progresivo desde **VoiceLDM base** (0,179), que tiene el valor más bajo, hasta **VoiceLDM + XTTS** (0,229), cuya barra es la más alta entre los modelos evaluados y la más cercana al **Ground Truth** (0,273). **Parler-TTS** (0,218) y **VoiceLDM fine-tuned** (0,213) presentan barras

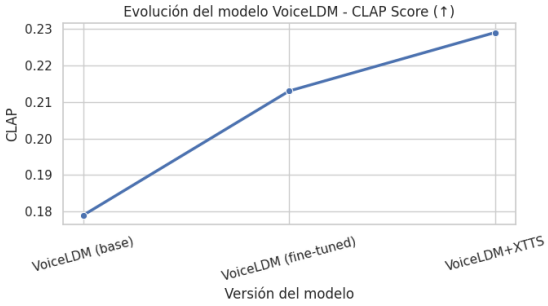
de altura similar, superando a **fairseq** (0,188), pero sin alcanzar el nivel del enfoque propuesto.

En el Gráfico de WER, se observa que el modelo **VoiceLDM base** tiene la barra más alta (47,9 %) y el modelo **VoiceLDM + XTTS**, presenta la barra más baja entre los modelos (18,5 %).

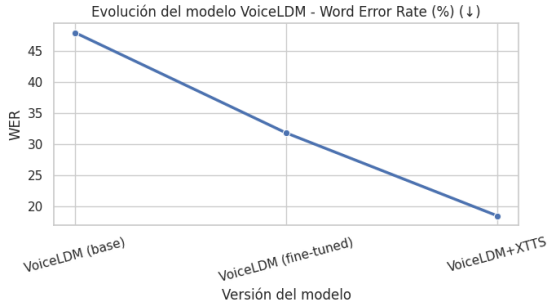
Para analizar el impacto del proceso de fine-tuning y la posterior integración con XTTS, en la Figura 5.2 se muestran gráficamente los valores de FAD, CLAP Score y WER en las diferentes etapas: modelo base, modelo ajustado y enfoque final VoiceLDM + XTTS.



((a)) Evolución del Modelo VoiceLDM - FAD ↓



((b)) Evolución del Modelo VoiceLDM - CLAP Score ↑



((c)) Evolución del Modelo VoiceLDM - WER ↓

Figura 5.2: Evolución del modelo VoiceLDM en distintas métricas tras el fine-tuning e integración con XTTS.

Los gráficos muestran mejoras consistentes en todas las métricas tras el ajuste fino y la combinación con XTTS. Específicamente, el FAD se reduce de 6,317 a 2,082, el FD de 7,723 a 1,406, el CLAP Score mejora de 0,179 a 0,229 y el WER se reduce de 47,9 % a 18,5 %.

Estos resultados validan la efectividad del enfoque propuesto, no solo por su capacidad para generar audio perceptivamente natural, sino también por su habilidad para preservar el contenido textual y semántico de la entrada de manera precisa y coherente.

5.5. Síntesis del Capítulo

En este capítulo se ha evaluado el rendimiento del enfoque propuesto para la generación de audio sintético multilingüe y multiacento, utilizando métricas cuantitativas objetivas y comparándolo directamente con modelos representativos del estado del arte. El fine-tuning del modelo VoiceLDM ha demostrado mejoras significativas en calidad acústica, inteligibilidad y coherencia contextual, evidenciadas por reducciones en FAD (de 6,317 a 5,261), FD (de 7,723 a 6,687) y WER (de 47,9 % a 31,8 %), así como un aumento en el CLAP Score (de 0,179 a 0,213). La integración con XTTS refuerza aún más el rendimiento, alcanzando un WER de 18,5 % y un CLAP Score de 0,229, superando a modelos como fairseq y Parler-TTS en precisión lingüística y fidelidad semántica. No obstante, Parler-TTS destaca en calidad acústica con un FAD de 1,406. En conjunto, las métricas cuantitativas (FAD, FD, KL, CLAP Score, WER) y las visualizaciones gráficas confirman la robustez del sistema propuesto.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones

En este trabajo se ha llevado a cabo el desarrollo de un marco innovador para la generación de audio sintético multilenguaje y multiacento que permite tener control de las características de la voz, como edad, género, idioma y acento. A través de la integración de los modelos VoiceLDM y XTTS, se ha propuesto una solución que supera las limitaciones de los modelos actuales, los cuales suelen estar restringidos en su capacidad para controlar simultáneamente múltiples atributos vocales o para generar audio en diversos idiomas con alta fidelidad. Los resultados experimentales obtenidos muestran que el enfoque propuesto logra un equilibrio sobresaliente entre calidad acústica, coherencia semántica y fidelidad lingüística, destacándose por su capacidad para generar audios personalizados que superan a modelos del estado del arte.

El proceso de fine-tuning de VoiceLDM utilizando conjuntos de datos de audio limpios (Common Voice 18.0, LibriTTS y CML-TTS) permitió mejorar notablemente su rendimiento, reflejado en la reducción de métricas como la FAD, que pasó de 6,317 a 5,261, y la FD, que disminuyó de 7,723 a 6,687. Además, el WER se redujo del 47,9 % al 31,8 %, mientras que el CLAP Score aumentó de 0,179 a 0,213, acercándose al valor del audio de referencia (0,273), lo que evidencia una mejora en la alineación semántica y la naturalidad de los audios generados.

La inclusión de un discriminador basado en Wav2Vec 2.0 permitió verificar automáticamente que los audios generados cumplen con las características deseadas de edad y género. Esta incorporación aseguró la coherencia entre las condiciones de entrada y las propiedades del audio generado, proporcionando al enfoque propuesto un mecanismo robusto de verificación interna y contribuyendo a su fiabilidad.

La integración de XTTS permitió adaptar los audios generados por VoiceLDM a distintos idiomas, superando así las limitaciones de idioma del modelo base, que solo genera audios en inglés, español y portugués, adaptando los audios a diferentes idiomas a la vez que preserva las características vocales del hablante. Los resultados evidencian que

la combinación VoiceLDM + XTTS alcanza una calidad destacada, logrando un WER de 18,5 %, el más bajo entre los modelos evaluados y un CLAP Score de 0,229, el más cercano al *ground truth*. Asimismo, se obtuvieron valores de FAD (2,082) y KL (0,003) muy próximos a los de los audios reales, lo que demuestra la capacidad del sistema para mantener la naturalidad y coherencia contextual en múltiples idiomas.

La evaluación comparativa con modelos del estado del arte, como fairseq y Parler-TTS, refuerza la ventaja del enfoque propuesto. Aunque Parler-TTS obtuvo mejores resultados en métricas de calidad acústica (FAD = 1,406 y FD = 0,584), el enfoque VoiceLDM + XTTS destacó en coherencia semántica (CLAP Score = 0,229) y precisión lingüística (WER = 18,5 %), superando a ambos modelos en estas dimensiones críticas. Esto sugiere que el sistema propuesto logra un equilibrio óptimo entre calidad acústica, fidelidad lingüística y alineación contextual.

Por último, el desarrollo de la API facilita gestionar las funcionalidades del marco propuesto, permitiendo su integración en aplicaciones reales, ofreciendo un control preciso sobre los atributos de la voz y una interfaz accesible para usuarios finales. No obstante, el rendimiento óptimo requiere entornos computacionales avanzados, lo que podría limitar su adopción en contextos con recursos restringidos. Este trabajo sienta las bases para futuras investigaciones que exploren la generalización a idiomas de baja disponibilidad de datos y la optimización de la eficiencia computacional, consolidando un paso adelante en la síntesis de audio sintético personalizable, multilingüe y multiacento.

6.2. Trabajo Futuro

Aunque el enfoque VoiceLDM + XTTS ha demostrado ser eficaz en la generación de audio sintético personalizado multilingüe y multiacento, aún existen desafíos importantes relacionados con la verificación automática y en tiempo real de la coherencia entre el audio generado y los atributos especificados (edad, género, idioma, acento). Aunque el discriminador utilizado para este trabajo es efectivo, su evaluación ocurre de manera posterior a la generación del audio, lo que puede introducir latencia y no garantiza una verificación en tiempo real de la alineación con los atributos deseados. Además, las métricas actuales, como el WER, FAD, KL y CLAP Score, resultan útiles para evaluar el rendimiento de forma externa y posterior, pero no permiten garantizar, durante el proceso de generación, que el audio mantenga la fidelidad con las características solicitadas. Para abordar esta limitación, se propone reemplazar el discriminador actual por un enfoque adversario basado en una red GAN, que actúe como un crítico integrado en el pipeline de generación.

6.2.1. Entrenamiento del Modelo Adversario

El modelo GAN constaría de dos componentes principales: un generador y un discriminador, entrenados de manera conjunta en un esquema adversario.

- **Generador**: Continuaría siendo el sistema actual (VoiceLDM + XTTS), responsable de generar audio sintético a partir de texto, condicionándolo con atributos como edad, género, idioma y acento. El generador se optimizaría para .[en]gañar.[a]l discriminador, produciendo audios que sean indistinguibles de los reales y que cumplan con las especificaciones de entrada.
- **Discriminador**: Se entrenaría para distinguir entre audios reales, extraídos de conjuntos de datos como Common Voice 18.0, LibriTTS y CML-TTS, y audios generados por el modelo. El objetivo del discriminador sería identificar incoherencias, como un acento incorrecto, un tono de voz que no corresponde al rango de edad especificado o un género vocal inconsistente con la entrada.

Este enfoque adversario tendría el potencial de mejorar la robustez del sistema al integrar la verificación de coherencia directamente en el proceso de generación, reduciendo la necesidad de iteraciones posteriores y mejorando la eficiencia en aplicaciones en tiempo real.

6.2.2. Otras Direcciones para el Trabajo Futuro

Además del enfoque adversario, se identifican varias áreas complementarias para extender el trabajo actual:

1. **Expansión de la cobertura lingüística y de acentos**: Aunque el sistema actual soporta múltiples idiomas, su cobertura se limita a los idiomas incluidos en los conjuntos de datos utilizados (inglés, español y portugués, con adaptaciones multilingües vía XTTS). En el futuro, se podría ampliar el entrenamiento con conjuntos de datos más diversos, para incluir idiomas menos representados, como dialectos regionales. Esto mejoraría la accesibilidad del sistema en contextos culturales diversos.
2. **Optimización para aplicaciones en tiempo real**: La latencia actual del sistema, especialmente en la fase de verificación con el discriminador, podría limitar su uso en aplicaciones en tiempo real, como asistentes virtuales interactivos. Se propone explorar técnicas de optimización, como la reducción de la complejidad de los modelos o el uso de arquitecturas más ligeras, para disminuir el tiempo de procesamiento sin comprometer la calidad.
3. **Evaluación subjetiva con usuarios reales**: Aunque las métricas objetivas (FAD, FD, WER, CLAP Score) proporcionan una evaluación robusta, la percepción humana de la calidad y naturalidad del audio es igualmente importante. En el futuro,

se podrían realizar pruebas subjetivas con hablantes nativos de diferentes idiomas y acentos para evaluar la aceptabilidad y autenticidad de los audios generados, complementando las métricas cuantitativas.

Bibliografía

[ABD+19] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[AR24] Hawraz A Ahmad and Tarik A Rashid. Planning the development of text-to-speech synthesis models and datasets with dynamic deep learning. *Journal of King Saud University-Computer and Information Sciences*, page 102131, 2024.

[Bet23] James Betker. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*, 2023.

[BH24] Matej Božić and Marko Horvat. A survey of deep learning audio generation methods. *arXiv preprint arXiv:2406.00146*, 2024.

[BMV+22] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation.(2022). *arXiv preprint arXiv:2209.03143*, 2022.

[BTS+24] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*, 2024.

[BWW+23] Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. Speech-based age and gender prediction with transformers. 2023.

[CDG+24] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*, 2024.

[CLL21] Hyunseung Chung, Sang-Hoon Lee, and Seong-Whan Lee. Reinforce-aligner: Reinforcement alignment search for robust end-to-end text-to-speech. *arXiv preprint arXiv:2106.02830*, 2021.

[CWS+22] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

[DCL+24] Tianjiao Du, Jun Chen, Jiasheng Lu, Qinmei Xu, Huan Liao, Yupeng Chen, and Zhiyong Wu. Controllable text-to-audio generation with training-free temporal guidance diffusion. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[GAD+17a] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, 30, 2017.

[GAD+17b] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, 30, 2017.

[GMMP23] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3590–3598, 2023.

[Has23] Mohammad Reza Hasanabadi. An overview of text-to-speech systems and media applications. *arXiv preprint arXiv:2310.14301*, 2023.

[HHY+23] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

[HJL+24] Yuhang He, Yash Jain, Xubo Liu, Andrew Markham, and Vibhav Vineet. Ritta: Modeling event relations in text-to-audio generation. *arXiv preprint arXiv:2412.15922*, 2024.

[HQR+17] Robert Hoehndorf, Núria Queralt-Rosinach, et al. Data science and symbolic ai: Synergies, challenges and opportunities. *Data Science*, 1(1-2):27–38, 2017.

[IBM23] IBM. Autoencoders variacionales (vae): Qué son y cómo funcionan. https://www.ibm.com/es-es/think/topics/variational-autoencoder, 2023. Última consulta: 19 de mayo de 2025.

[IBM24] IBM. ¿qué es un modelo generativo? https://www.ibm.com/es-es/think/topics/generative-model, 2024. Última consulta: 19 de mayo de 2025.

[KCI+20] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[KKS23] Yogesh Kumar, Apeksha Koul, and Chamkaur Singh. A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. *Multimedia Tools and Applications*, 82(10):15171–15197, 2023.

[KPK+23] Gokul Karthik Kumar, SV Praveen, Pratyush Kumar, Mitesh M Khapra, and Karthik Nandakumar. Towards building text-to-speech systems for the next billion users. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[KS23] Navdeep Kaur and Parminder Singh. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, 56(7):5837–5880, 2023.

[KSP+22] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

[KVB+23] Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718, 2023.

[KZK+23] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*, 2023.

[Łań21] Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE, 2021.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[LCY+23] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[LDY13] Zhen-Hua Ling, Li Deng, and Dong Yu. Modeling spectral envelopes using restricted boltzmann machines for statistical parametric speech synthesis. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7825–7829. IEEE, 2013.

[LGS+] Yichong Leng, ZHifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, Kaitao Song, Lei He, et al. Prompttts 2: Describing and generating voices with text prompt. In *The Twelfth International Conference on Learning Representations*.

[LK24] Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*, 2024.

[LYNC24] Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung. Voiceldm: Text-to-speech with environmental context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12566–12571. IEEE, 2024.

[MRVPK23] V Madhusudhana Reddy, T Vaishnavi, and K Pavan Kumar. Speech-to-text and text-to-speech recognition using deep learning. In *2023 2nd International Conference on Edge Computing and Applications (ICECAA), IEEE Xplore, Namakkal*, 2023.

[NM21] Owais Nazir and Aruna Malik. Deep learning end to end speech synthesis: A review. In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, pages 66–71. IEEE, 2021.

[OCJ+23] Frederico S Oliveira, Edresson Casanova, Arnaldo Candido Junior, Anderson S Soares, and Arlindo R Galvão Filho. Cml-tts: A multilingual dataset for speech synthesis in low-resource languages. In *International Conference on Text, Speech, and Dialogue*, pages 188–199. Springer, 2023.

[Pab24] Pablo Huet. Técnicas clave para el procesamiento de texto en nlp. https://openwebinars.net/blog/tecnicas-clave-para-procesamiento-texto-nlp/, July 2024. Última consulta: 19 de mayo de 2025.

[PPG+18] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *International Conference on Learning Representations*, 2018.

[RRT+19] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

[Sch22] Juergen Schmidhuber. Annotated history of modern ai and deep learning. *arXiv preprint arXiv:2212.11279*, 2022.

[Ser22] Sertis. Latent diffusion models: A review (part i). https://sertiscorp.medium.com/latent-diffusion-models-a-review-part-i-d0feacc4906, December 2022. Última consulta: 19 de mayo de 2025.

[Sha24] Somesh Sharma. Benefits or concerns of ai: A multistakeholder responsibility. *Futures*, 157:103328, 2024.

[SMK+17] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

[SNT24] Surabhi Sudhan, Parvathy P Nair, and Mg Thushara. Text-to-speech and speech-to-text models: A systematic examination of diverse approaches. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pages 1–8. IEEE, 2024.

[Spe22] Speechify. What is an autoregressive voice model? https://speechify.com/blog/autoregressive-voice-model/, September 2022. Última consulta: 19 de mayo de 2025.

[Sum24] Sumit Singh. Enhancing text-to-audio multimodal systems: Fine-tuning, evaluation metrics, and real-world applications. https://www.labellerr.com/blog/enhancing-text-to-audio-multimodal-systems-fine-tuning-evaluation-metrics-and-real-world-applications/, July 2024. Última consulta: 19 de mayo de 2025.

[Syr25] Syracuse University. What is machine learning? key concepts and real-world uses. https://ischool.syracuse.edu/what-is-machine-learning/, February 2025. Última consulta: 19 de mayo de 2025.

[TQSL21] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

[TSİ+23] Andreas Triantafyllopoulos, Björn W Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, et al. An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*, 111(10):1355–1381, 2023.

[TUA18] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4784–4788. IEEE, 2018.

[WCM+21] Andrew Werchniak, Roberto Barra Chicote, Yuriy Mishchenko, Jasha Droppo, Jeff Condal, Peng Liu, and Anish Shah. Exploring the application of synthetic audio in training keyword spotters. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7993–7996, 2021.

[WCZ+23] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[WHA+21] Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. fairseq: A scalable and integrable speech synthesis toolkit. *arXiv preprint arXiv:2109.06912*, 2021.

[WSRS+17] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[WTT24] Zixuan Wang, Yu-Wing Tai, and Chi-Keung Tang. Audio-agent: Leveraging llms for audio generation, editing and composition. *arXiv preprint arXiv:2410.03335*, 2024.

[XDGL24] Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *arXiv preprint arXiv:2401.01044*, 2024.

[YLL+24] Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Retrieval-augmented text-to-audio generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585. IEEE, 2024.

[YYW+23] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.

[ZBW20] Shuai Zhao, Frede Blaabjerg, and Huai Wang. An overview of artificial intelligence applications for power electronics. *IEEE Transactions on Power Electronics*, 36(4):4633–4658, 2020.

[Zil25] Zilliz. Getting started with audio data: Processing techniques and key challenges. https://medium.com/@zilliz_learn/getting-started-with-audio-data-processing-techniques-and-key-challenges-420dc5233163, February 2025. Última consulta: 10 de marzo de 2025.

Anexos

Apéndice A

Análisis Exploratorio

En esta sección, se presenta un análisis exploratorio de datos (EDA, por sus siglas en inglés) del conjunto de datos de entrenamiento utilizado para el desarrollo del modelo VoiceLDM. Este conjunto, derivado de Common Voice, LibriTTS y CML-TTS, contiene grabaciones de audios en inglés, español y portugués. Incluye metadatos asociados a cada audio, como identificadores de hablantes, transcripciones, edad, género, acento, duración e idioma. El análisis exploratorio permite caracterizar la distribución de estas variables, identificar patrones y evaluar la calidad y representatividad del conjunto de datos, aspectos fundamentales para garantizar un entrenamiento robusto y equilibrado del modelo.

A.1. Resumen del Dataset

A.1.1. Dimensiones del Conjunto de Datos

El conjunto de datos de entrenamiento contiene un total de 1.201.771 muestras distribuidas en 10 columnas, incluyendo `speaker_id`, `file_path`, `text`, `age`, `gender`, `accent`, `description`, `duration`, `language` y `text_length`. En la Tabla A.1 se puede observar un resumen de esto.

Cuadro A.1: Dimensiones del conjunto de datos de entrenamiento utilizado para VoiceLDM

Característica	**Valor**
Número de muestras	1,201,771
Número de columnas	10

A.1.2. Revisión de Valores Nulos y Tipos de Datos

La revisión de tipos de datos y la presencia de valores nulos es un paso fundamental para asegurar la calidad del conjunto de datos antes del entrenamiento. En este caso, se verificó que las 1.201.771 muestras contienen valores no nulos en todas las columnas. Además, los tipos de datos son coherentes con el contenido esperado de cada variable: la

duración del audio se encuentra en formato numérico decimal (float64), la longitud del texto como entero (int64), y el resto de columnas como cadenas de texto (object). En la Tabla A.2 se presenta un resumen de esto.

Cuadro A.2: Conteo de valores nulos y tipo de dato por columna

Columna	Valores no nulos	Tipo de dato
speaker_id	1,201,771	object
file_path	1,201,771	object
text	1,201,771	object
age	1,201,771	object
gender	1,201,771	object
accent	1,201,771	object
description	1,201,771	object
duration	1,201,771	float64
language	1,201,771	object
text_length	1,201,771	int64

A.2. Análisis de Variables Categóricas

El análisis de las variables categóricas del conjunto de datos permite identificar el equilibrio de las muestras en relación con los factores demográficos y lingüísticos que influyen en la generación de voz sintética. Se consideran las variables age, gender, language y accent, todas ellas relevantes para garantizar diversidad y representatividad en el entrenamiento del modelo VoiceLDM. Este análisis incluye tanto visualizaciones de frecuencia como tablas resumen con proporciones, lo que permite una interpretación precisa de la composición del conjunto de datos.

A.2.1. Distribución de la Edad

La variable age describe el rango de edad de cada hablante. La Figura A.1 muestra una clara concentración de hablantes en los rangos de edad de los veinte, treinta y cuarenta años, que representan más del 70 % del total. Los rangos de edad extremos como niños, ochentas y noventas están muy poco representadas. La Tabla A.3 detalla la distribución absoluta y relativa por categoría.

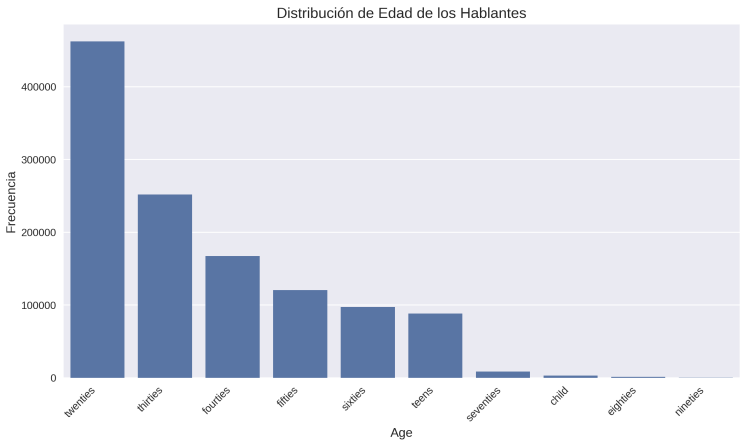


Figura A.1: Distribución de edad de los hablantes.

Cuadro A.3: Distribución de la edad de los hablantes

Categoría	**Frecuencia**	**Porcentaje (%)**
twenties	462 492	38,48
thirties	252 027	20,97
fourties	167 382	13,93
fifties	120 476	10,02
sixties	97 484	8,11
teens	88 367	7,35
seventies	8 685	0,72
child	3 209	0,27
eighties	1 497	0,12
nineties	152	0,01

A.2.2. Distribución del Género

La distribución por género presenta un desequilibrio notable, con una mayoría de hablantes masculinos (74.2 %). La Figura A.2 representa gráficamente esta proporción, mientras que la Tabla A.4 recoge los valores exactos.

Cuadro A.4: Distribución de género de los hablantes

Categoría	**Frecuencia**	**Porcentaje (%)**
male	891 690	74,2
female	310 081	25,8

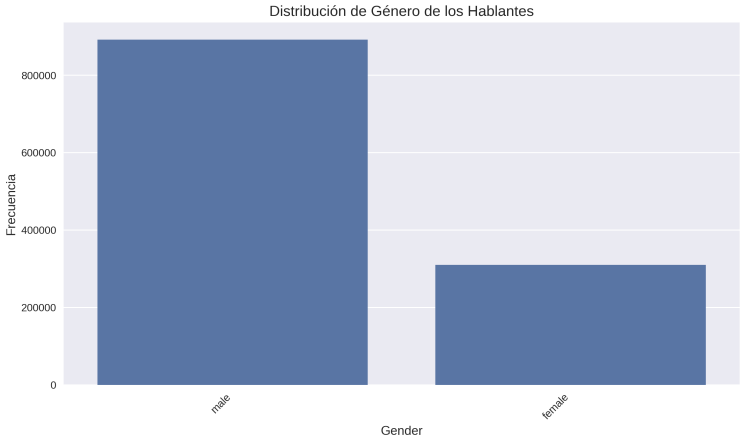


Figura A.2: Distribución del Género de los Hablantes.

A.2.3. Distribución del Idioma

Como se observa en la Figura A.3 y la Tabla A.5, el idioma predominante en el conjunto de datos es el inglés (74.74 %), seguido por el español (23.43 %) y en menor medida por el portugués (1.82 %). Esta distribución refleja la disponibilidad pública de datos y la composición original de los conjuntos de datos fuente (Common Voice, LibriTTS y CML-TTS).

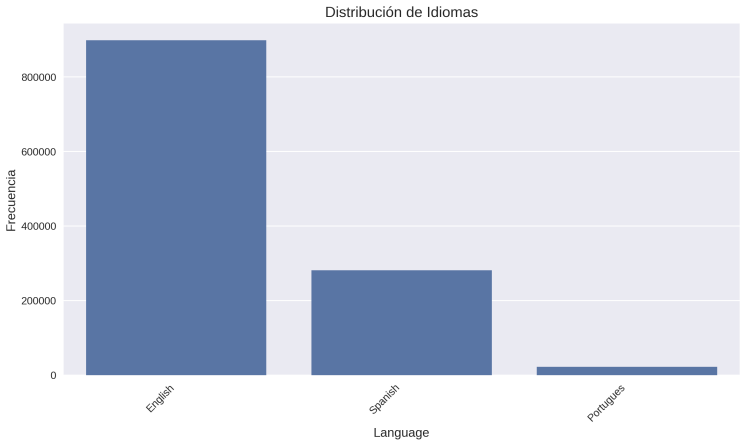


Figura A.3: Distribución del Idioma de los Hablantes.

Cuadro A.5: Distribución de idiomas en el conjunto de datos

Categoría	**Frecuencia**	**Porcentaje (%)**
English	898 247	74,74
Spanish	281 598	23,43
Portuguese	21 926	1,82

A.2.4. Distribución del Acentos

La variable accent presenta una alta diversidad, pero en esta sección se analizan los diez acentos más representados. La Figura A.4 y la Tabla A.6 muestran que los acentos más frecuentes son el inglés americano, el inglés británico y el español mexicano-estadounidense.

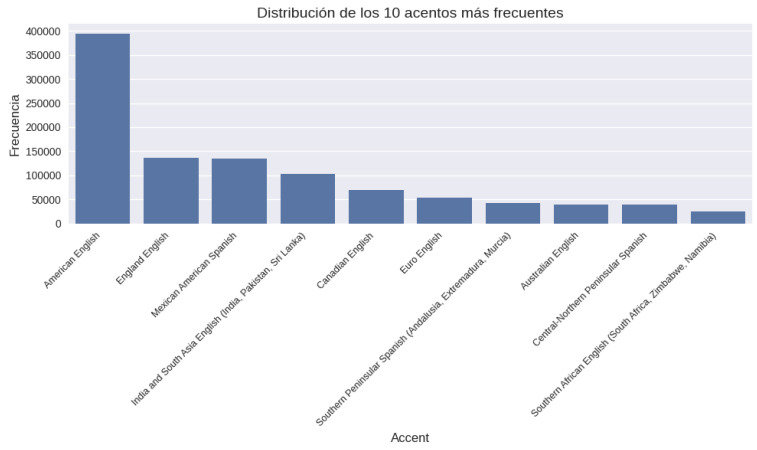


Figura A.4: Distribución de los 10 Acentos más Frecuentes.

Cuadro A.6: Distribución de los 10 acentos más frecuentes

Categoría	**Frecuencia**	**Porcentaje (%)**
American English	394 271	32,81
England English	136 821	11,38
Mexican American Spanish	135 227	11,25
India and South Asia English	102 693	8,55
Canadian English	68 959	5,74
Euro English	53 794	4,48
Southern Peninsular Spanish	43 132	3,59
Australian English	40 144	3,34
Central-Northern Peninsular Spanish	38 967	3,24
Southern African English	25 909	2,16

A.3. Análisis de Variables Continuas

El análisis de variables continuas permite explorar características numéricas fundamentales del conjunto de datos. En esta sección, se examina en detalle la duración de los segmentos de audio como variable fundamental en la generación de voz.

A.3.1. Duración de los Audios

La variable duration representa la duración de cada segmento de audio en segundos. La Figura A.5 muestra la distribución de duraciones en el conjunto de datos, evidenciando una forma aproximadamente normal con una concentración en torno a los 5 segundos.

El análisis estadístico detallado, presentado en la Tabla A.7, indica que la mayoría de los audios tienen una duración entre 3,82 y 6,16 segundos (rango intercuartílico), con una duración mínima de 0,25 segundos y una máxima de casi 15 segundos. Esta amplitud en las duraciones proporciona al modelo una mayor exposición a muestras de diferente longitud, favoreciendo la generalización y la robustez del sistema de generación.

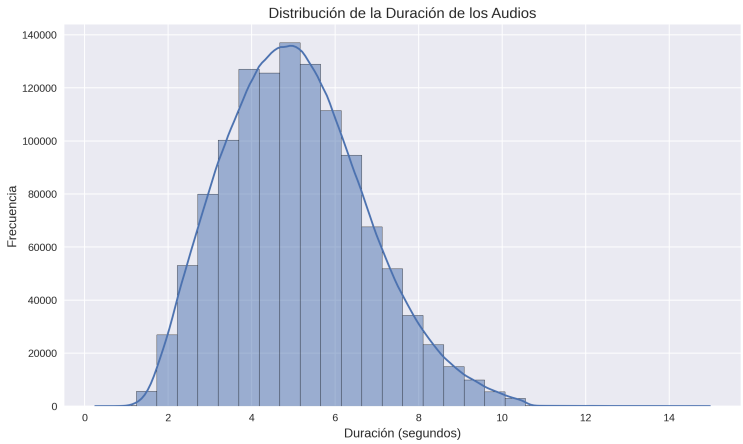


Figura A.5: Distribución de la Duración de los Audios.

Cuadro A.7: Estadísticas descriptivas de la duración de los audios (en segundos)

Estadística	**Valor**
Cantidad de muestras	1 201 771
Media	5,06
Desviación estándar	1,68
Mínimo	0,25
Percentil 25 (Q1)	3,82
Mediana (Q2)	4,97
Percentil 75 (Q3)	6,16
Máximo	14,98

A.4. Relaciones entre Variables

En esta sección se exploran las relaciones existentes entre diferentes variables del conjunto de datos para identificar patrones y asociaciones relevantes. Analizar cómo se interrelacionan variables categóricas y continuas, como género, edad, idioma, acento y

duración de los audios, permite comprender mejor la composición y diversidad del conjunto de datos.

A.4.1. Distribución de Género por Idioma

La distribución de género por idioma, representada en la Figura A.6, muestra la proporción relativa de hablantes masculinos y femeninos dentro de cada idioma del conjunto de datos. Como se observa, el inglés presenta una mayoría predominante de hablantes masculinos, seguida por el español y el portugués. Esta distribución porcentual, detallada en la Tabla A.8, evidencia cómo la variable género varía según el idioma.

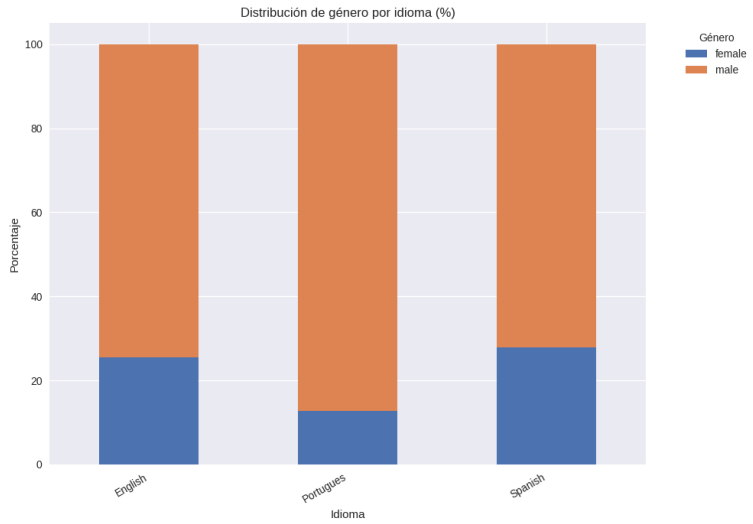


Figura A.6: Distribución de Género por Idioma (%).

Cuadro A.8: Distribución de Género por Idioma (%)

Idioma	**Femenino**	**Masculino**
English	25,47	74,53
Portuguese	12,70	87,30
Spanish	27,88	72,12

A.4.2. Edad vs Acento

Se analiza la relación entre las variables categóricas `accent` y `edad` para evaluar la diversidad demográfica del conjunto de datos y su representatividad en términos de edad para diferentes regiones lingüísticas. La Figura A.7 muestra un gráfico de barras apiladas que ilustra la distribución porcentual de los rangos de edad en los 10 acentos más frecuentes. Puede observarse, por ejemplo, que los acentos asociados a regiones de habla inglesa como *American English* y *England English* presentan una distribución de rangos de edad más equilibrada, mientras que otros, como *Mexican American Spanish* y *India and South Asia English*, están fuertemente concentrados en el grupo de edad *twenties*.

La Tabla A.9 complementa esta visualización mostrando los porcentajes exactos por rangos de edad y acento, lo que permite observar de forma más precisa la distribución demográfica.

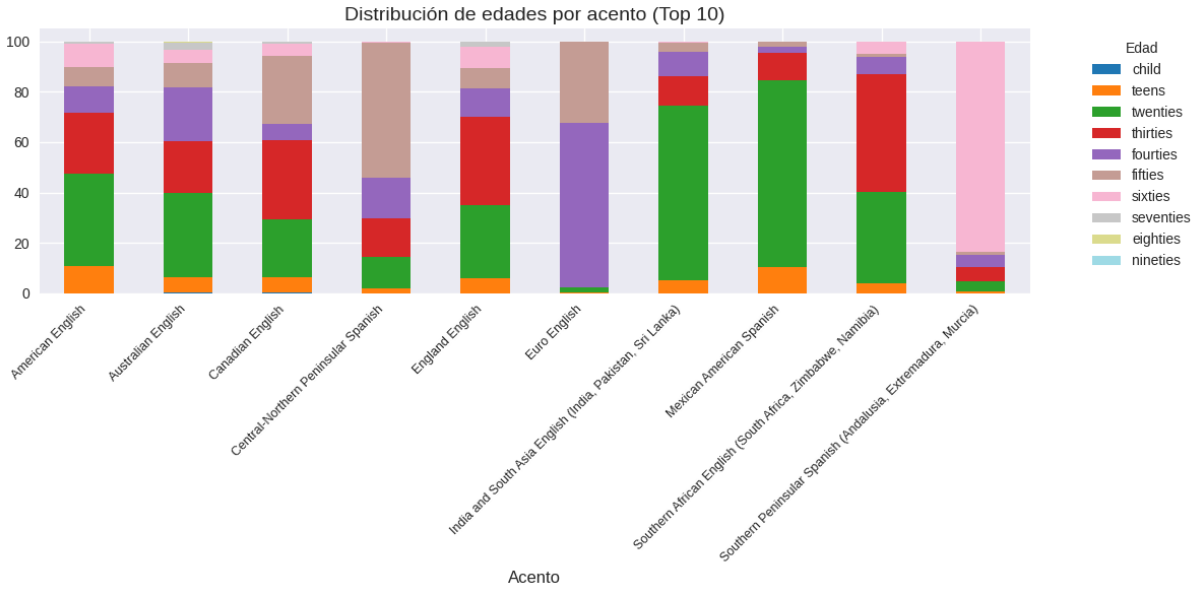


Figura A.7: Distribución de Edades por Acento (Top 10).

Cuadro A.9: Distribución de edades por acento (Top 10) (%)

Acento	**Child**	**Teens**	**Twenties**	**Thirties**	**Fourties**	**Fifties**	**Sixties**	**Seventies**	**Eighties**	**Nineties**
American English	0,07	10,66	36,73	24,39	10,16	7,81	9,13	0,82	0,22	0,01
Australian English	0,28	6,13	33,59	20,57	21,01	9,92	5,27	2,61	0,63	0,00
Canadian English	0,31	6,16	22,95	31,21	6,79	26,68	4,96	0,91	0,03	0,00
Central-Northern Peninsular Spanish	0,00	1,88	12,48	15,38	16,19	53,48	0,52	0,07	0,00	0,00
England English	0,00	5,91	28,99	35,24	11,15	8,03	8,43	2,00	0,17	0,07
Euro English	0,00	0,40	1,91	0,15	65,37	32,17	0,00	0,00	0,00	0,00
India and South Asia English	0,00	5,15	69,37	11,67	9,68	3,79	0,33	0,00	0,00	0,00
Mexican American Spanish	0,00	10,45	74,25	10,73	2,25	2,19	0,07	0,05	0,00	0,00
Southern African English	0,00	3,98	36,32	46,74	6,70	1,38	4,73	0,15	0,00	0,00
Southern Peninsular Spanish	0,00	0,83	3,80	5,94	4,87	1,17	83,40	0,00	0,00	0,00

A.4.3. Duración de los Audios por Género

Se analiza la relación entre la duración de los audios y el género de los hablantes para identificar posibles diferencias en la duración de las grabaciones entre hombres y mujeres. La Figura A.8 muestra un diagrama de caja (boxplot) que compara la distribución de la variable continua `duration` según la variable categórica `gender`.

Como se observa en la figura, las distribuciones de duración son similares en forma y dispersión para ambos géneros. Sin embargo, los valores centrales muestran ligeras diferencias. Las mujeres presentan una duración media de 5,23 segundos, mientras que los hombres alcanzan una media de 5,01 segundos. Además, el rango intercuartílico es ligeramente superior en el grupo femenino, lo que indica una mayor variabilidad. Estos resultados se resumen en la Tabla A.10, que incluye estadísticas descriptivas como la media, desviación estándar, cuartiles y valores extremos.

Este análisis sugiere que, aunque la duración promedio de los audios es relativamente consistente entre géneros, existe una leve tendencia a duraciones ligeramente mayores en las muestras correspondientes a mujeres.

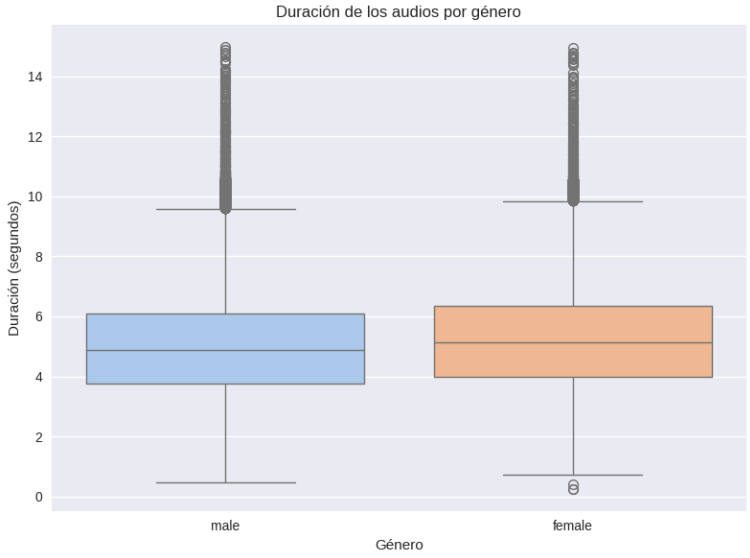


Figura A.8: Distribución de la duración de los audios por género.

Cuadro A.10: Estadísticas descriptivas de la duración por género (en segundos)

Género	N	Media	Desv. Std.	Mín.	Q1	Mediana	Q3	Máx.
Femenino	310081	5,23	1,68	0,25	4,00	5,15	6,34	14,95
Masculino	891690	5,01	1,68	0,47	3,77	4,90	6,10	14,98

A.4.4. Distribución de Género por Acento

Se analiza la distribución de género en función del acento, utilizando los diez acentos más representados en el conjunto de datos. El objetivo es observar si existen desequilibrios de género dentro de cada variedad lingüística y entender mejor la composición demográfica de los datos.

La Figura A.9 muestra un gráfico de barras apiladas que representa el porcentaje relativo de hablantes masculinos y femeninos para cada acento. Se observan contrastes importantes: por ejemplo, el acento *Southern African English* presenta una mayoría femenina (64,84 %), mientras que en *Euro English* la proporción masculina es prácticamente absoluta (99,95 %). También destacan acentos más equilibrados, como *Mexican American Spanish* y *India and South Asia English*, con distribuciones de género relativamente más simétricas.

La Tabla A.11 ofrece los valores porcentuales exactos para cada combinación de acento y género, lo que permite un análisis cuantitativo más preciso.

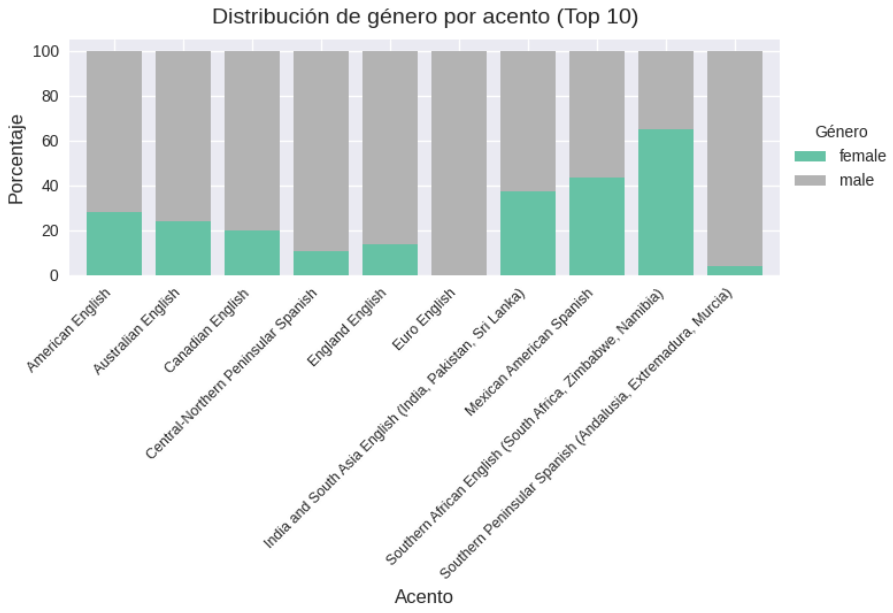


Figura A.9: Distribución de Género por Acento.

Cuadro A.11: Distribución de Género por Acento (%)

Acento	**Femenino**	**Masculino**
American English	28,02	71,98
Australian English	23,87	76,13
Canadian English	19,76	80,24
Central-Northern Peninsular Spanish	10,56	89,44
England English	13,78	86,22
Euro English	0,05	99,95
India and South Asia English	37,48	62,52
Mexican American Spanish	43,60	56,40
Southern African English	64,84	35,16
Southern Peninsular Spanish	3,82	96,18

Apéndice B

API para la Gestión de las Funcionalidades del Marco Propuesto

Se ha desarrollado una API que ofrece un conjunto completo de funcionalidades para la generación y gestión de audios sintéticos. Estas funcionalidades se agrupan en diferentes endpoints, cada uno con un propósito específico, como se describe a continuación:

B.0.1. Consultar las Combinaciones de los Atributos de Voz Disponibles

Este endpoint permite recuperar todas las combinaciones disponibles de atributos de voz que pueden ser utilizadas para la generación de audio sintético. Sirve como punto de consulta previa para conocer las configuraciones válidas soportadas por el sistema, facilitando la construcción de peticiones correctas.

El endpoint retorna una lista de objetos JSON. Cada objeto representa una configuración específica de voz compatible con el generador de audio, especificando los siguientes campos:

- **language:** Idioma de la voz (“en”, “es”,“pt”).
- **accent:** Acento del idioma (por ejemplo, “United States English”, “England English”, “Central American Spanish”, “Rio de Janeiro Brazilian Portuguese”).
- **gender:** Género asociado a la voz (“male_masculine”,“female_feminine”).
- **age_category:** Grupo de edad aproximado de la voz (“child”, “teens”, “twenties”, “thirties”, “forties”, “fifties”, “sixties”, “eighties”, “nineties”).

Método HTTP: `GET`
Parámetros de entrada: Este endpoint no requiere parámetros de entrada. Al ser una

operación `GET`, simplemente devuelve los datos disponibles en el sistema relacionados con voces preconfiguradas.

Salida: Una estructura JSON con una lista de todas las combinaciones disponibles. Como se ilustra en la Figura B.1.

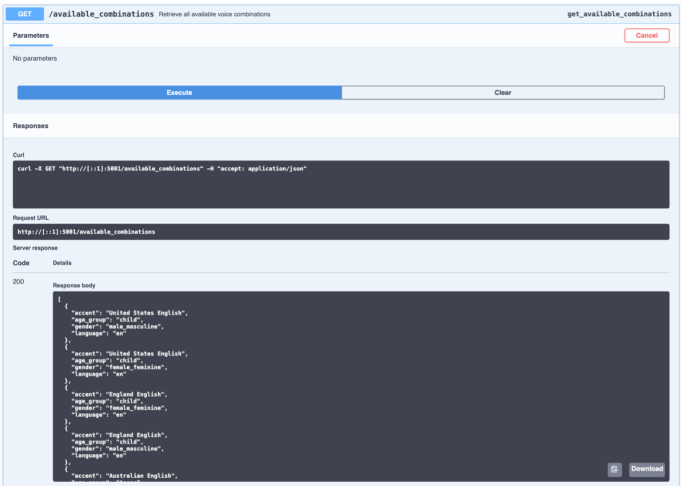


Figura B.1: Lista de todas las combinaciones de los atributos de la voz disponibles.

B.0.2. Generar una Nueva Voz

Este endpoint permite generar una nueva muestra de audio a partir de una entrada de texto y una configuración de atributos de voz seleccionada por el usuario. La operación se realiza de forma asincrónica: tras recibir la solicitud, el sistema pone en cola la tarea de generación y devuelve un *task_id* que puede utilizarse para consultar posteriormente el estado y resultado del procesamiento. La ejecución en segundo plano permite que el endpoint responda rápidamente sin bloquear el flujo de solicitudes.

Método HTTP: `POST`

Parámetros de entrada: Los parámetros deben enviarse como parte de la cadena de consulta (`query string`). Todos los campos son obligatorios.

- `text_prompt` (string): Contenido lingüístico del mensaje que se desea sintetizar en voz. Ejemplo: `"Hola, ¿cómo estás?"`
- `language` (string): Código del idioma. Valores válidos: “en”, “es”,“pt”.
- `accent` (string): Acento o variante regional de la voz. Ejemplo: “United States English”, “Spain Spanish”.
- `gender` (string): Género de la voz. Valores válidos: “female_feminine”, “male_masculine”.

- `age_group` (string): Grupo de edad de la voz. Valores válidos: “child”, “teens”, “twenties”, “thirties”, “forties”, “fifties”, “sixties”, “eighties”, “nineties”.

Funcionamiento interno:

- El sistema valida que todos los campos requeridos estén presentes.
- Se realiza una preprocesamiento del texto de entrada conforme al idioma seleccionado.
- Se genera un identificador único para la tarea de generación de voz (`task_id`).
- La tarea es colocada en una cola asincrónica para ser procesada en segundo plano, incluyendo.

Respuesta: En caso de éxito, se retorna un objeto JSON con el `task_id` asignado. Como se observa en la Figura B.2.

El ID generado sigue el formato: `<voice_id>_<audio_number>` (por ejemplo, `268bee7-1644-4d44-8e36-0fa31cdbaf9c_0`).

Este identificador puede usarse en el Endpoint B.0.4 para consultar el estado de la tarea y acceder al archivo de audio resultante.

Códigos de estado posibles:

- `200 OK`: La tarea fue registrada correctamente y se devuelve el `task_id`.
- `400 Bad Request`: Faltan parámetros requeridos o están mal formateados.
- `404 Not Found`: No se encontró una combinación de atributos de voz que coincida con los parámetros solicitados.
- `500 Internal Server Error`: Ocurrió un error interno durante el procesamiento o inicialización de la tarea.

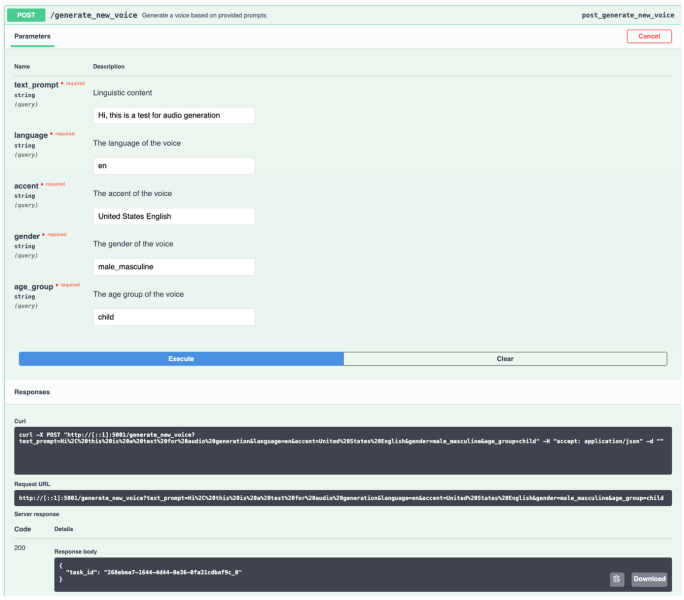


Figura B.2: Generar una nueva voz con los atributos y el contenido lingüístico especificado.

B.0.3. Subir un audio de referencia

Este endpoint permite a los usuarios cargar guna grabación de voz existente, con el fin de clonar una voz específica. A través de una solicitud `POST`, los usuarios pueden subir uno o más archivos de audio junto con información contextual de la voz (idioma, acento, género y grupo de edad). Una vez procesada exitosamente la carga, se genera un `voice_id` único que puede ser utilizado para tareas posteriores de generación de audio.

Método HTTP: `POST`

Entrada (multipart/form-data):

- **audios (array[.wav])**: Lista de archivos de audio en formato binario (preferentemente `.wav`) representando la voz que se desea clonar.
- **language (string)**: Idioma de la voz. Ejemplos: `en`, `es`, `pt`.
- **accent (string)**: Acento del hablante, como por ejemplo `United States English`, `Brazilian Portuguese`, etc.
- **gender (string)**: Género del hablante. Valores aceptados: `female_feminine`, `male_masculine`.
- **age_group (string)**: Grupo etario del hablante. Valores posibles: `teens`, `twenties`, `thirties`, `forties`, `fifties`, `sixties`, `eighties`, `nineties`.

Salida:

- `voice_id (string)`: Identificador único asignado a la voz cargada. Como se muestra en la Figura B.3.

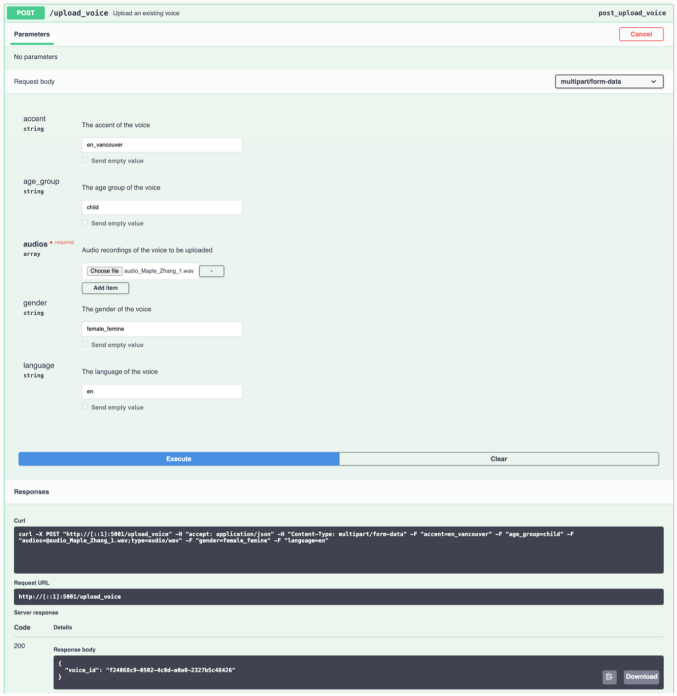


Figura B.3: Guardar un archivo de audio existente.

B.0.4. Consultar el Estado de una Tarea de Generación

Este endpoint permite consultar el estado de una tarea asincrónica iniciada previamente para la generación o clonación de audios. Al proporcionar un `task_id` válido como parámetro de consulta, el sistema responderá con el estado actual de la tarea. Esto permite a los usuarios saber si su solicitud aún está en cola, en proceso o ya ha finalizado. La Figura B.4 muestra un ejemplo de esto.

Método HTTP: `GET`

Parámetros de entrada:

- **task_id (string)**: Identificador único de la tarea cuya información se desea consultar.

Salida:

- `status (string)`: Estado actual de la tarea. Puede tomar uno de los siguientes valores:
  - `Pending`
  - `In progress`
  - `Completed`
- `voice_id (string)`: Identificador de la voz asociada a la tarea. Se incluye sólo si la tarea ha finalizado correctamente.

- `audio_number (string, opcional)`: Número de audio generado dentro del conjunto de audios para la voz correspondiente. Se incluye sólo si la tarea ha finalizado.

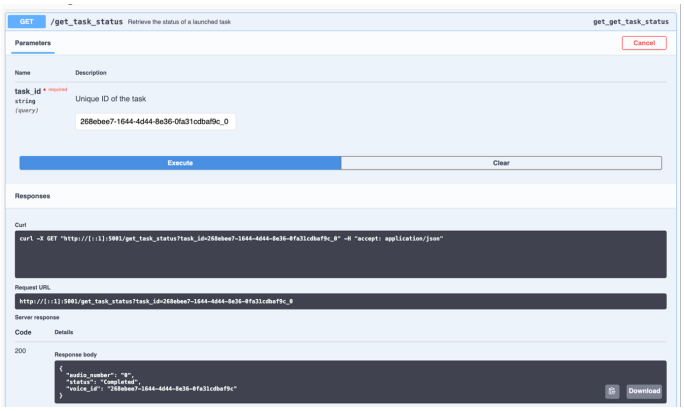


Figura B.4: Consultar el estado de la generación del audio.

B.0.5. Obtener un Audio en Específico

Este endpoint permite a los usuarios descargar un archivo de audio generado previamente, utilizando el identificador de voz (`voice_id`) y el número del audio (`audio_number`). La respuesta será un archivo en formato `.wav` si la solicitud es válida.

Método HTTP: `GET`

Parámetros de entrada:

- **voice_id (string)**: Identificador único de la voz previamente generada o cargada.
- **audio_number (string)**: Número del audio específico asociado al `voice_id`.

Salida:

- Archivo de audio en formato `audio/wav`, que se descarga automáticamente con el nombre `<voice_id>_<audio_number>.wav`. Como en el ejemplo de la Figura B.5.

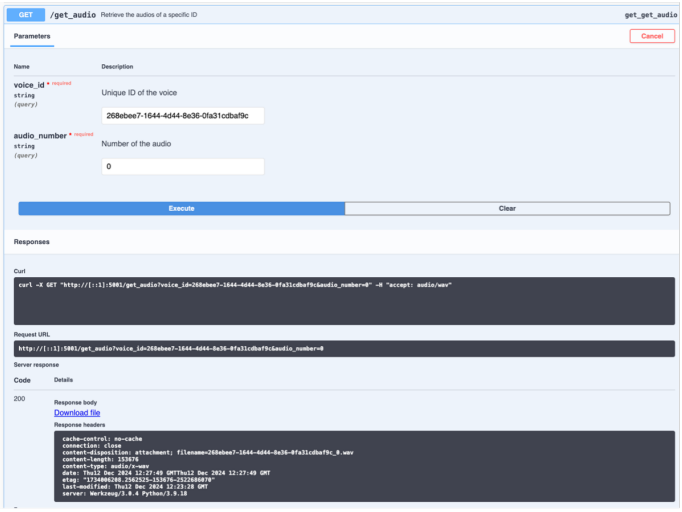


Figura B.5: Descargar el archivo de audio con el ID de la voz y el número de audio.

B.0.6. Obtener Todos los Archivos de Audio asociados a un ID de Voz Específico

Este endpoint permite a los usuarios descargar todos los archivos de audio generados o cargados bajo un identificador de voz específico (`voice_id`). La respuesta consiste en un archivo comprimido en formato ZIP que contiene todos los archivos `.wav` asociados a dicho ID. Como se muestra en la Figura B.6.

Método HTTP: `GET`

Parámetros de entrada:

- **voice_id (string)**: Identificador único de la voz. Es utilizado para recuperar todos los audios generados con dicha voz.

Salida:

- Archivo `.zip` que contiene todos los archivos de audio en formato `.wav` relacionados con el `voice_id` proporcionado.

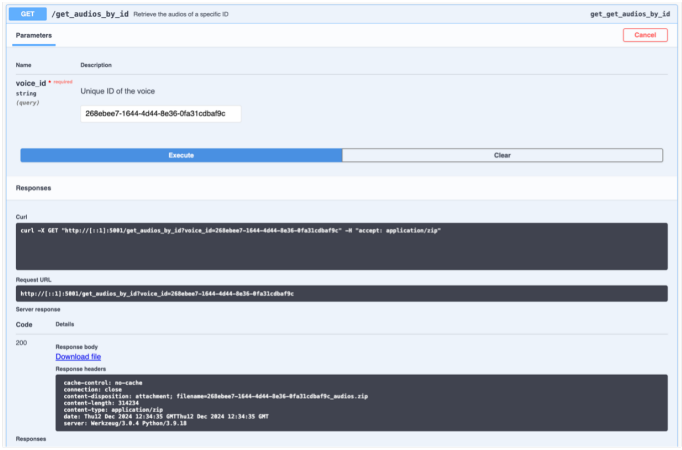


Figura B.6: Descargar todos los archivos de audio asociados a un ID de voz específico.

B.0.7. Listar Todos los Identificadores de Voz

Este endpoint permite obtener una lista de todos los identificadores de voz (`voice_id`) previamente generados, junto con información descriptiva asociada a cada voz. Es útil para obtener un resumen del conjunto de voces disponibles en el sistema. La Figura B.7 muestra un ejemplo de esto.

Método HTTP: `GET`

Parámetros de entrada: No se requiere ningún parámetro.

Salida: Una estructura JSON con la lista de todos los Identificadores de Voz Disponibles.

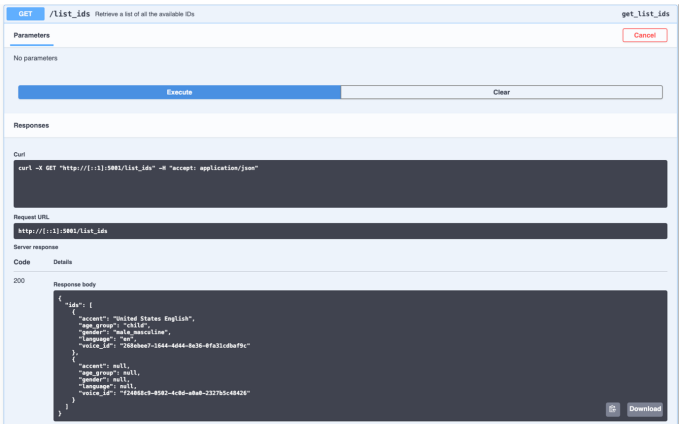


Figura B.7: Lista de los identificadores de voz de todas las voces generadas.

B.0.8. Clonar una Voz

Una vez que se tiene el ID de la voz deseada, es posible clonar dicha voz para producir un nuevo audio utilizando un texto de entrada. Este endpoint permite iniciar la generación de un nuevo audio basado en el `voice_id` especificado, un texto (`text_prompt`) y el idioma del contenido. La Figura B.8 muestra el proceso de clonación de una voz dado el ID de la voz deseada.
Método HTTP: `POST`

Parámetros de entrada (query):

- `voice_id` (string, requerido): Identificador único de la voz previamente generada.
- `text_prompt` (string, requerido): Texto lingüístico que se convertirá en audio.
- `language` (string, requerido): Idioma del texto de entrada. Valores aceptados: `en`, `es`, `pt`.

Respuesta: En caso de éxito, se retorna un objeto JSON con el `task_id` asignado.

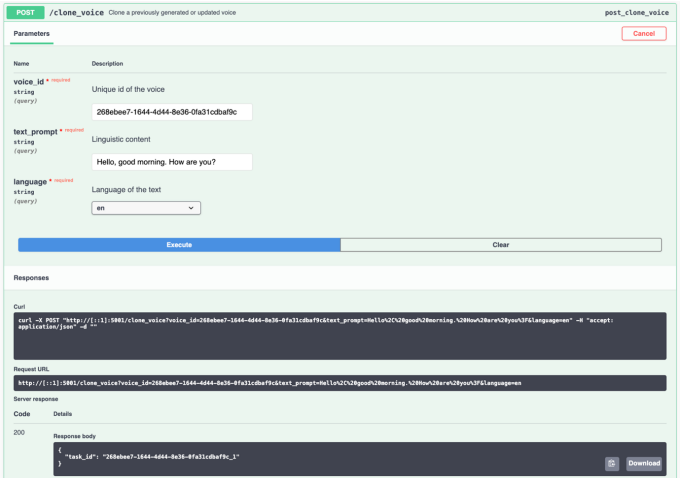


Figura B.8: Clonar una voz existente.

B.0.9. Eliminar un Archivo de Audio Específico

Este endpoint permite eliminar un archivo de audio específico asociado a una voz previamente generada. Para llevar a cabo la eliminación, el usuario debe proporcionar el `voice_id` y el número del audio (`audio_number`). En la Figura B.9 se muestra un ejemplo del proceso para eliminar un audio.

Método HTTP: `DELETE`
Parámetros de entrada:

- `voice_id` (string): Identificador único de la voz.

- `audio_number` (string): Número del archivo de audio que se desea eliminar (por ejemplo, `audio2` si el archivo es `audio2.wav`).

Respuesta: En caso de éxito, se retorna un mensaje de confirmación de la eliminación del archivo de audio especificado.

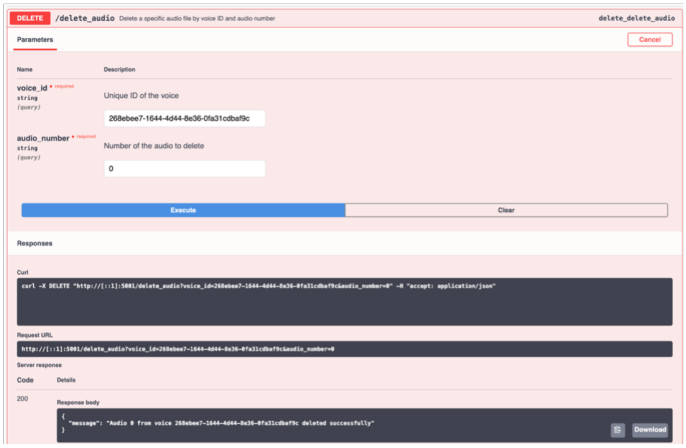


Figura B.9: Eliminar un archivo de audio por el ID de la voz y el número del audio.

B.0.10. Eliminar una Voz

Este endpoint permite eliminar completamente una voz del sistema utilizando su identificador único (`voice_id`). Esta operación elimina de forma permanente todos los audios asociados a esa voz, así como sus metadatos. En la Figura B.10 se muestra un ejemplo para eliminar una voz.

Método HTTP: `DELETE`

Parámetros de entrada:

- `voice_id` (string): Identificador único de la voz que se desea eliminar.

Salida: Si la operación es exitosa, se devuelve un mensaje confirmando la eliminación de la voz.

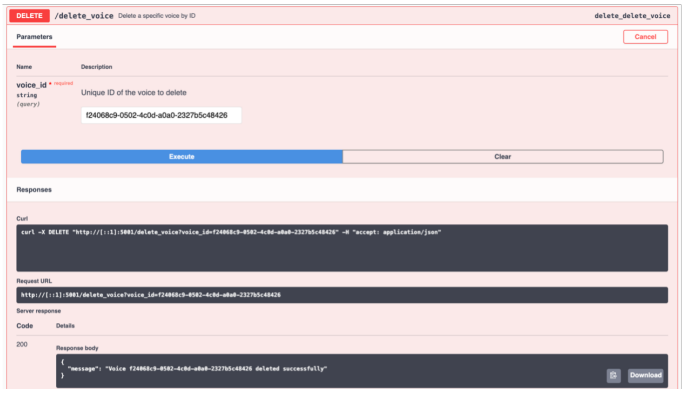


Figura B.10: Eliminar una voz específica por el ID.

Apéndice C

Artículo Científico Derivado del TFM

Como resultado de este Trabajo de Fin de Máster, se redactó y envió un artículo científico titulado:

Advanced Synthetic Audio Generator: A Comprehensive Multilingual and Multi-accent Approach

El artículo está siendo revisado por la revista *IEEE Transactions on Multimedia*, una revista de alto impacto dedicada a la investigación en tecnología multimedia y sus aplicaciones, abarcando áreas como procesamiento de señales, sistemas, software e integración de sistemas.

En este manuscrito se sintetizan los principales aspectos abordados en el TFM, incluyendo:

- La propuesta metodológica y la arquitectura basada en los modelos VoiceLDM y XTTS.
- El enfoque multilingüe y multiacento para la generación de audio sintético.
- El proceso de fine-tuning con datos limpios y diversos.
- La incorporación de un discriminador para validar atributos del hablante como edad y género.
- La evaluación cuantitativa con métricas como FAD, FD, KL, CLAP Score y WER.
- La comparación con modelos del estado del arte y el análisis de resultados.

Este envío representa un logro significativo en el trabajo desarrollado, ya que contribuye a la difusión científica del conocimiento generado y refuerza la validez académica y técnica del enfoque propuesto.