UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE CUENCA

CARRERA DE COMPUTACIÓN

DISEÑO Y DESARROLLO DE UN MÓDULO DE CONVERSIÓN DE TEXTO A VOZ EN BASE A TÉCNICAS DE APRENDIZAJE PROFUNDO PARA EL ASISTENTE ROBÓTICO VOLK DE LA CÁTEDRA UNESCO DE LA UPS

Trabajo de titulación previo a la obtención del
título de Ingeniero en Ciencias de la Computación

AUTORES: KAAR JOSEPH MASHINGASHI UNKUCH
JUAN SEBASTIÁN ANDRADE ALULEMA
TUTOR: ING. CRISTIAN FERNANDO TIMBI SISALIMA

Cuenca - Ecuador
2025

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Kaar Joseph Mashingashi Unkuch con documento de identificación N° 1400730618 y Juan Sebastián Andrade Alulema con documento de identificación N° 0302448642; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Cuenca, 17 de febrero del 2025

Atentamente,

Kaar Joseph Mashingashi Unkuch	Juan Sebastián Andrade Alulema
1400730618	0302448642

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotros, Kaar Joseph Mashingashi Unkuch con documento de identificación N° 1400730618 y Juan Sebastián Andrade Alulema con documento de identificación N° 0302448642, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Proyecto técnico: "Diseño y desarrollo de un módulo de conversión de texto a voz en base a técnicas de aprendizaje profundo para el asistente robótico Volk de la Cátedra UNESCO de la UPS", el cual ha sido desarrollado para optar por el título de: Ingeniero en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, 17 de febrero del 2025

Atentamente,

Kaar Joseph Mashingashi Unkuch	Juan Sebastián Andrade Alulema
1400730618	0302448642

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Cristian Fernando Timbi Sisalima con documento de identificación N° 0103709911, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: DISEÑO Y DESARROLLO DE UN MÓDULO DE CONVERSIÓN DE TEXTO A VOZ EN BASE A TÉCNICAS DE APRENDIZAJE PROFUNDO PARA EL ASISTENTE ROBÓTICO VOLK DE LA CÁTEDRA UNESCO DE LA UPS, realizado por Kaar Joseph Mashingashi Unkuch con documento de identificación N° 1400730618 y por Juan Sebastián Andrade Alulema con documento de identificación N° 0302448642, obteniendo como resultado final el trabajo de titulación bajo la opción Proyecto técnico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Cuenca, 17 de febrero del 2025

Atentamente,

Ing. Cristian Fernando Timbi Sisalima
0103709911

DEDICATORIA

Dedico este trabajo académico con todo mi corazón a las personas que han iluminado mi camino con amor, sacrificio y sabiduría. A Dios, mi guía y refugio, quien ha sido mi fuerza en los momentos de incertidumbre, iluminando mi camino y recordándome que, incluso en las adversidades, nunca estoy solo.

A mi madre, Lidsay, quien personifica el amor incondicional y la fortaleza. Su entrega, sacrificios y fe en mí han sido el motor que me impulsa. Este logro es tanto suyo como mío, porque sin su apoyo, nada de esto sería posible.

A mi hermano, Meset, mi compañero de vida. Su apoyo constante ha sido fundamental en mis momentos de alegría y dificultad. Su presencia me inspira y me da fuerzas para seguir adelante.

A mis abuelos, Verónica y Tomás, quienes con su ejemplo me enseñaron a vivir con humildad, esfuerzo y fe. Ellos me han dado las herramientas para construir mis sueños y su legado sigue vivo en cada paso que doy.

A mis tías, Sayda y Kenia, por su cariño constante y apoyo incondicional, y a mi tío Chitup, cuyo ejemplo sigue siendo fuente de inspiración. A mi querido tío Kevin, cuya partida dejó un vacío, pero cuyo amor y enseñanzas siguen vivos en mi corazón.

A todos ustedes, dedico este logro, porque sin su amor y apoyo, este sueño no habría sido posible. Este trabajo es un tributo a cada uno de ustedes que me ha dado tanto, y lo llevo en mi corazón con gratitud eterna.

Kaar Joseph Mashingashi Unkuch

DEDICATORIA

Dedico este proyecto de titulación a Dios Todopoderoso, ya que ha sido mi fuente infinita de amor, fortaleza y sabiduría. Sin su guía, su acompañamiento día y noche, sus fuerzas y su inmenso amor hacia mí, nada de esto hubiera sido posible. Le entrego cada uno de mis logros y sueños cumplidos y por cumplir; en todo momento fue, es y será mi luz y mi refugio.

A mi madre, Nancy Pamela Alulema Dávila, quien con su gran amor e incondicional ha sido mi roca desde que tengo memoria. Su apoyo constante, cada uno de sus sacrificios por mi y hacia mi y su fe puesta en mí me han impulsado a alcanzar cada meta que me he propuesto.

A mis dos padres, Byron Teodoro Andrade Aguirre y Juan Sebastián Ochoa Ordoñez, ejemplos de excelencia profesional y valores inquebrantables. Son mi gran inspiración, al enseñarme con su vida lo que significa trabajar duro y jamás rendirse.

A mi hermana, Martina Sofía Ochoa Alulema, quien ha sido mi mayor motivo para dar siempre lo mejor de mí. Todo lo que hago es para ser un ejemplo digno para ella, demostrando que con esfuerzo y perseverancia cada sueño se puede cumplir.

A mi primer y gran amor, mi mejor amiga incondicional, Ana Isabel Baculima Durán. Su compañía inquebrantable, su apoyo constante, su amor y su fe, por su fortaleza puesta en mí para cada uno de los momentos buenos y malos, fáciles y difíciles, han sido pilares fundamentales para no rendirme, por ser mi fuente de inspiración de ser una gran estudiante y fantástica persona, el cual su amor y apoyo me ha hecho seguir adelante en este increíble proceso hacia mi profesionalización.

Gracias a todos ustedes, siempre les tendré mi gratitud, mi amor y mi dedicación, porque fueron, son y siempre serán la razón detrás de cada paso que doy y cada sueño que cumpla.

Juan Sebastian Andrade Alulema

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a la Universidad Politécnica Salesiana, por brindarme las herramientas y el espacio para crecer no solo como profesional, sino también como persona. A los docentes, quienes con dedicación y pasión compartieron sus conocimientos, dejando en mí una huella que perdurará para siempre. Su compromiso con la educación ha sido un faro que ha iluminado mi camino.

De manera especial, quiero agradecer a mi tutor de tesis, quien no solo me brindó su conocimiento, sino también su tiempo, paciencia y apoyo incondicional. Su guía fue fundamental para transformar este proyecto en una realidad. Gracias por creer en mí, incluso cuando yo dudaba de mí mismo.

A mis amigos y compañeros de estudio, quienes con su compañerismo, empatía y alegría hicieron de este camino una experiencia inolvidable. Las risas, los debates, las noches de estudio y los momentos de desahogo son recuerdos que atesoro profundamente. Su apoyo hizo que los retos fueran más llevaderos.

A mi familia, especialmente a mi madre, Lidsay, mi hermano, Meset, y mis abuelos, quienes siempre me alentaron a seguir adelante. A cada uno de ellos, por su amor incondicional, paciencia y fe en mí, incluso en los momentos de dificultad. Este logro es también suyo, porque sin su apoyo constante, nunca habría llegado hasta aquí.

Finalmente, a todas aquellas personas que, de manera directa o indirecta, han contribuido a que este sueño se haga realidad. Este logro no es solo mío; es el resultado de un esfuerzo colectivo, de la confianza de quienes creyeron en mí y del amor de quienes me han acompañado en cada paso.

Kaar Joseph Mashingashi Unkuch

AGRADECIMIENTOS

En primer lugar, agradezco a Dios Todopoderoso, quien ha sido mi guía y refugio en cada momento de este proceso. Su amor y fortaleza me han sostenido en los momentos más duros y difíciles, dándome así la fuerza que necesito, la perseverancia y la fe necesarias para alcanzar esta meta.

A mi familia, porque su amor incondicional, paciencia y apoyo constante han sido la base sobre la cual he construido este sueño. A mi madre, mis padres y mi hermana, por creer en mí incluso cuando yo dudaba. Su confianza, sacrificio y enseñanzas me han dado las herramientas para enfrentar cada desafío.

A Ana Isabel Baculima Durán, mi gran compañera de vida y mi mejor amiga, la cual me ha apoyado desde el primer momento en el que nos conocimos, por estar a mi lado en todo momento, dándome fuerzas cuando más las necesitaba y celebrando cada uno de mis logros como si fueran suyos. Su amor, apoyo y fe en mí han sido fundamentales para llegar hasta aquí.

A mis profesores y mentores, por compartir su conocimiento, por guiarme con paciencia y exigencia, y por inspirarme a seguir aprendiendo cada día. A cada uno de ellos, mi profundo respeto y gratitud por haber sido parte de mi formación académica y profesional.

A mi universidad, por brindarme un espacio de aprendizaje, crecimiento y oportunidades. A la comunidad universitaria que me permitió desarrollarme no solo como profesional, sino también como persona.

A todos mis grandes amigos, compañeros y personas que de alguna manera siempre han dejado y dejarán su huella en este camino y en mi corazón. Por su compañía, su apoyo y por hacer de esta experiencia algo inolvidable.

Gracias a todos ustedes. Siempre les tendré mi más grande gratitud, mi amor y mi dedicación, los llevaré a cada uno de ustedes dentro de mi corazón, porque fueron, son y siempre serán la razón detrás de cada paso que doy y cada sueño que cumplo.

Juan Sebastián Andrade Alulema

Resumen

Este proyecto tiene como objetivo el diseño y desarrollo de un módulo de conversión de texto a voz (TTS) para el asistente robótico Volk, creado en la Cátedra UNESCO Tecnologías de Apoyo para la Inclusión Educativa de la Universidad Politécnica Salesiana (UPS). La meta fue clonar una voz que sea clara, natural y empática, adecuada para aplicaciones educativas y personalizadas. Para ello, se emplean técnicas avanzadas de redes neuronales profundas para generar una voz fluida, adaptable y de alta calidad. A diferencia de los métodos tradicionales, que dependen de fragmentos pregrabados, este enfoque permite la clonación de voz, ofreciendo una identidad vocal única, coherente y empática para el asistente robótico Volk.

La metodología utilizada para guiar el análisis de datos en este módulo es CRISP-DM, que sigue un enfoque estructurado dividido en varias fases: análisis de datos, preparación, modelado, evaluación e implementación. El desarrollo del módulo, por otro lado, se gestionó mediante la metodología Scrum, que permitió una ejecución ágil y colaborativa, asegurando que cada etapa del proceso contribuya de manera significativa a la creación de un sistema eficiente, funcional y alineado con los requisitos técnicos y las expectativas de los usuarios.

Este sistema está orientado a trabajar en español de Latinoamérica, adaptando la prosodia y las características lingüísticas necesarias para una síntesis de voz precisa y culturalmente relevante. El resultado es una solución accesible y eficiente que mejora la interacción de los usuarios con la tecnología, cumpliendo con los objetivos de naturalidad y empatía requeridos en diversas aplicaciones.

Palabras clave: conversión de texto a voz (TTS), asistente robótico, CRISP-DM, machine learning, Scrum.

Abstract

This project aims to design and develop a text-to-speech (TTS) module for the Volk robotic assistant, created at the UNESCO Chair in Assistive Technologies for Educational Inclusion at the Salesian Polytechnic University (UPS). The goal was to clone a voice that is clear, natural and empathetic, suitable for educational and personalized applications. To do this, advanced deep neural network techniques are employed to generate a fluid, adaptive and high quality voice. Unlike traditional methods, which rely on pre-recorded fragments, this approach enables voice cloning, providing a unique, coherent and empathetic voice identity for the Volk robotic assistant.

The methodology used to guide the data analysis in this module is CRISP-DM, which follows a structured approach divided into several phases: data analysis, preparation, modeling, evaluation and implementation. The development of the module, on the other hand, was managed using the Scrum methodology, which allowed an agile and collaborative execution, ensuring that each stage of the process contributes significantly to the creation of an efficient, functional system aligned with technical requirements and user expectations.

This system is oriented to work in Latin American Spanish, adapting the prosody and linguistic characteristics necessary for an accurate and culturally relevant speech synthesis. The result is an accessible and efficient solution that improves user interaction with technology, meeting the objectives of naturalness and empathy required in various applications.

Keys words: text-to-speech (TTS), robotic assistant, CRISP-DM, machine learning, Scrum.

ÍNDICE DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

I. **Introducción**.

La conversión de texto a voz (TTS) es una tecnología fundamental para mejorar la interacción entre las personas y los sistemas automatizados, permitiendo una comunicación más natural y efectiva. El asistente robótico Volk, desarrollado por la Universidad Politécnica Salesiana (UPS), busca integrar una voz sintética que, además de ser comprensible y funcional, cumpla con los estándares de naturalidad y empatía necesarios para aplicaciones en entornos educativos y asistencias personalizadas. Este proyecto tiene como objetivo el diseño y desarrollo de dicho sistema, adaptado a las características y necesidades específicas del usuario.

Es así, que este documento aborda los fundamentos y el proceso de desarrollo del sistema, comenzando con el análisis del problema y los objetivos generales y específicos. Se profundiza en aspectos teóricos clave, como la conversión de texto a voz, la importancia de la prosodia en TTS y el uso de un *dataset* propio en español de Latinoamérica. También se detalla el enfoque metodológico adoptado, que combina el modelo CRISP-DM para la gestión de datos y la metodología Scrum para la gestión iterativa del desarrollo, con la inclusión de tecnologías como redes neuronales recurrentes, transformadores y procesamiento del lenguaje natural (PLN).

Seguidamente, para el desarrollo se incluyó la selección de herramientas, la configuración del entorno de trabajo y la recopilación de datos necesarios para entrenar el modelo TTS. Además, se implementó un sistema ajustado a las características prosódicas del español de Latinoamérica, considerando factores como la entonación, el ritmo y la modulación de la voz.

Este documento tiene la estructura que se indica a continuación: la **Sección II** aborda el problema que motiva este trabajo, la **Sección III** presenta los objetivos generales y específicos, y la **Sección IV** detalla los fundamentos teóricos, incluyendo la conversión de texto a voz, la prosodia en TTS y los fundamentos tecnológicos. El Marco Metodológico se encuentra en la **Sección V**, mientras que los resultados, cronograma, presupuesto, conclusiones y recomendaciones se desarrollan en las **Secciones VI a IX**, complementadas por referencias y anexos en las Secciones **X y XI**.

II. **Problema.**

El desarrollo de interfaces naturales y fluidas entre humanos y máquinas es uno de los mayores desafíos en la inteligencia artificial, especialmente en el campo de los asistentes robóticos (Morales et al., 2021). A pesar de los avances tecnológicos en la conversión de texto a voz (TTS), persisten limitaciones en la capacidad de generar voces con entonación y prosodia adecuadas. Estos aspectos son cruciales para aproximar la interacción con los robots a una conversación verdaderamente humana, lo cual es esencial para mejorar la efectividad de la comunicación y la satisfacción del usuario. El asistente robótico Volk, desarrollado por la Universidad Politécnica Salesiana, requiere un sistema de TTS que no solo sea comprensible y

funcional, sino que también posea una voz que imite la modulación humana, enfatizando palabras, pausando adecuadamente y expresando emociones. Actualmente, este desafío afecta principalmente a robots educativos y asistivos, limitando su efectividad y aceptación.

La percepción de los usuarios sobre la calidad de la interacción con los asistentes robóticos depende en gran medida de la calidad de la voz generada. La voz de los asistentes debe ir más allá de ser simplemente inteligible; debe ser natural y expresar la misma fluidez que una conversación entre personas (González & López, 2022). La dificultad radica en generar voces que logren este nivel de naturalidad, ya que las tecnologías actuales de TTS enfrentan desafíos en cuanto a la entonación y los acentos adecuados. Esto impacta negativamente la confianza del usuario y la adopción de estos sistemas, especialmente en entornos donde la comunicación efectiva es esencial, como en la educación y la asistencia personalizada (Gómez et al., 2023). Además, la falta de naturalidad también puede desmotivar a los usuarios a interactuar con los robots, lo que limita su aplicación en campos donde la empatía y la comprensión son cruciales, como en la asistencia a personas mayores o con necesidades especiales.

Con respecto al impacto potencial de este proyecto se extiende más allá del asistente Volk. Al mejorar la conversión de texto a voz, se puede generar una comunicación más efectiva y empática, lo que aumentará la adopción de estas tecnologías en diversos ámbitos. Un sistema de voz más natural y comprensible también facilita la integración de los robots en entornos educativos y de asistencia diaria; áreas en las que la interacción humana es fundamental para una experiencia exitosa. Además, la capacidad de los asistentes robóticos para interactuar de manera más humana y empática contribuirá a una mayor aceptación de la tecnología, mejorando la experiencia general y promoviendo el uso de estos sistemas en aplicaciones especializadas (Hansen, 2023).

Un sistema de voz que permita una interacción más natural no solo mejoraría la calidad de la comunicación, sino que también facilitaría el uso de estos asistentes en entornos donde la empatía y la comprensión son clave, como la educación o la asistencia a personas con necesidades especiales. De esta manera, mejorar la voz del asistente Volk contribuiría a crear un ambiente en el que los usuarios se sientan más cómodos y confiados al interactuar con la tecnología, promoviendo así una mayor adopción de estos sistemas (Pérez, 2024). Los avances logrados en este proyecto tienen el potencial de beneficiar a un público mucho más amplio y fomentar el uso de tecnologías asistivas en diversos sectores.

III. Objetivos Generales y Específicos.

Objetivo General

Diseñar y desarrollar un módulo de conversión de texto a voz basado en técnicas de aprendizaje profundo que integre entonación y prosodia, con el fin de mejorar la calidad de la interacción entre usuarios y el asistente robótico Volk, evaluando su efectividad mediante pruebas comparativas y la percepción de los usuarios.

Objetivos Específicos

OE1: Estudiar y evaluar las principales redes neuronales de aprendizaje profundo y herramientas para la generación de texto a voz, considerando aspectos de entonación y prosodia, mediante la realización de pruebas comparativas.

OE2: Desarrollar un módulo que permita realizar la conversión de texto a voz para generación de diálogos empleando entonación y prosodia para el asistente robótico.

OE3: Ejecutar un plan de experimentación que permita medir el nivel de percepción que tienen los usuarios en relación al módulo de conversión de texto a voz.

IV. Revisión de la literatura o fundamentos teóricos.

La revisión de la literatura que sustenta el diseño y desarrollo de un módulo de conversión de texto a voz en base a técnicas de aprendizaje profundo para el asistente robótico volk de la cátedra Unesco de la UPS. Se parte de la Conversión de Texto a Voz (TTS), dónde se plantea la evolución de los TTS. Seguidamente, se conocerá la importancia de la prosodia en TTS. Además, se abordará sobre el Dataset Propio en Español de Latinoamérica, y su diversidad para el proceso de recolección de datos que nos facilitará el desarrollo de este proyecto. También, es importante una revisión teórica de la Metodología CRISP-DM; mismo que nos guiará al Entendimiento del Negocio para el Análisis de los Datos como la Preparación de los Datos, el modelado, la evaluación y el despliegue. Considerando su importancia, el Método de aplicación se analizará la teoría que nos guiará para la fase de la práctica. Así mismo, en los Fundamentos Tecnológicos es importante analizar la teoría para comprender las Redes Neuronales Recurrentes (RNN) y Transformadores, Frecuencias Agradables y Empatía en la Voz Sintética, Procesamiento del Lenguaje Natural (PLN). Y finalmente, los aportes teóricos de las Perspectivas Futuras y Tecnologías Emergentes en la Síntesis de Voz que nos dará cumplimiento a los objetivos de este proyecto.

4.1 Conversión de texto a voz (TTS)

La transformación de texto en voz (TTS) hace referencia al proceso de transformar texto escrito en una señal de voz sintetizada. Esta tecnología ha evolucionado significativamente, pasando de los primeros enfoques basados en la concatenación de fragmentos pregrabados, a sistemas que utilizan modelos avanzados de redes neuronales profundas, así como plantean los autores Camero & Mejía (2021); que la transformación de texto a voz (TTS) es una técnica madura para la cual existen múltiples herramientas capaces de generar voz en tiempo real.

Además, la Evolución de los sistemas TTS hace referencia que en los sistemas iniciales, la síntesis de voz se lograba mediante la concatenación de unidades pregrabadas, lo que limitaba la flexibilidad y la naturalidad del habla generada (Aquilué, 2024). Con la introducción de redes neuronales profundas, se lograron avances notables, especialmente con modelos como Tacotron

2 y WaveNet, que revolucionaron la síntesis de voz al generar señales acústicas desde representaciones textuales de manera más natural y precisa (Salgado, 2024). Estos avances permitieron una mayor fluidez y naturalidad en la generación de la voz, aunque no logran captar completamente la complejidad de la prosodia empática en el contexto del español de Latinoamérica.

Para este proyecto, se destaca la necesidad de desarrollar un *dataset* propio, específico para el dialecto del español de Latinoamérica, lo cual permitirá entrenar un modelo que refleje mejor las características prosódicas y emocionales del lenguaje.

Con respecto a la prosodia, se refiere a los elementos suprasegmentales del habla, como la entonación, el ritmo y el énfasis, que juegan un rol importante en la percepción de la naturalidad y emoción en la voz sintetizada. Además, los sistemas TTS avanzados no solo deben reproducir el contenido del texto, sino también transmitir la intención emocional a través de la prosodia (Padilla, 2022).

4.2 Importancia de la prosodia en TTS

Un sistema TTS capaz de generar prosodia adecuada resulta fundamental para una interacción fluida y natural entre el usuario y el asistente robótico. Diversos estudios han demostrado que una prosodia pobre reduce significativamente la percepción de empatía en sistemas de voz artificial, lo que afecta la interacción humano-máquina (Gómez et al., 2023). En el caso del asistente robótico Volk, se requiere que la voz no solo sea comprensible, sino que también sea percibida como empática, cercana y capaz de ajustarse a diferentes contextos.

La implementación de prosodia empática en los sistemas TTS ha sido objeto de investigaciones recientes, dado su impacto en la aceptación por parte de los usuarios (King & Karaiskos, 2019). Este enfoque no solo se centra en generar una voz con correcta entonación, sino también en ajustar parámetros como la duración de las sílabas, la modulación tonal y el énfasis para transmitir emociones. Además, el objetivo para este proyecto se centra en incorporar estas características prosódicas en la voz generada para el asistente robótico Volk, mejorando así la interacción emocional entre el robot y los usuarios.

4.3 *Dataset* propio en español de Latinoamérica

Una de las etapas críticas de este proyecto es la creación de un *dataset* en español de Latinoamérica, específicamente diseñado para capturar las variaciones prosódicas y emocionales que existen en los distintos dialectos de América Latina.

4.3.1 Diversidad del español de Latinoamérica

El español de Latinoamérica es una lengua rica en variaciones dialectales que afectan tanto la entonación como el ritmo del habla. Este fenómeno está bien documentado y subraya la necesidad de desarrollar un sistema TTS que pueda adaptarse a estas variaciones (González &

López, 2022). La mayoría de los datasets actuales de TTS en español se centran en el español de España (La Real Academia Española - RAE), lo que no es adecuado para este proyecto, dado que el asistente robótico Volk está diseñado para usuarios de habla hispana en América Latina.

4.3.2 Proceso de recolección de datos

El proceso de recolección del *dataset* implica la recolección de audios de una persona hablante nativa de América Latina, lo que garantiza la representación de una amplia gama de acentos y variaciones prosódicas. Este enfoque asegura que el modelo TTS entrenado sea adaptable y pueda generar una voz que sea tanto comprensible como natural para los usuarios en diferentes países (González & López, 2022).

4.4 Metodología CRISP-DM

La metodología Cross-Industry Standard Process for Data Mining o CRISP-DM las cuales son sus siglas; ha sido un estándar considerablemente realizado para y en proyectos de análisis y minería de datos. Misma, que fue desarrollada en el año 90 por un consorcio de empresas para crear una estructura flexible y escalable; con la finalidad de realizar análisis de datos en diversas industrias. Este modelo ha ganado popularidad debido a su carácter cíclico, que permite realizar iteraciones continuas a lo largo del proyecto (Wirth & Hipp, 2000).

Además, se considera importante integrar la metodología CRISP-DM; misma que versa de seis fases principales que cubren todas las fases del desarrollo de un proyecto o trabajo para análisis de datos, desde el entendimiento inicial de los problemas, hasta la realización de la solución definitiva (Sánchez & Pérez, 2023).

4.4.1 Entendimiento del negocio

Este punto es el más importante y el objetivo principal, para así comprender los requerimientos y objetivos del negocio. Se busca saber y poder identificar los problemas clave que se desean resolver mediante el análisis de datos. Cada uno de los objetivos del negocio se traducen en problemas que se pueden abordar mediante técnicas de minería de datos. Esta transformación es crucial, ya que guía todo el proceso de análisis (Torres &, 2021). De esta manera, la minería de datos ha dado lugar a una paulatina sustitución del análisis de datos por un enfoque de análisis de datos (Jaramillo & Arias, 2015).

4.4.2 Análisis de los datos

Después de establecer los objetivos del negocio, se pone en marcha la recolección y el análisis de los datos disponibles, dónde esta etapa consiste en conocer a fondo las características de los datos, su fiabilidad y posibles problemas. Aquí se evalúan cuestiones como la existencia de valores faltantes o atípicos, y se realizan análisis iniciales para conseguir una perspectiva general del conjunto de datos (Larose & Larose, 2014). Además, esta técnica es útil en

investigaciones donde los datos faltantes son escasos y en el caso de variables continuas o discretas sencillas nos facilita en análisis (Gómez, G. H., 2024).

4.4.3 Preparación de los datos

El autor Ferrero (2024), plantea que la fase de análisis es una etapa importante en el proceso de minería de datos; luego de transformar los datos se utilizan diversas técnicas para encontrar patrones y relaciones en los datos. Además, la preparación de los datos incluye la limpieza, transformación y estructuración de los datos para que puedan ser utilizados en los modelos de análisis. Este es un proceso laborioso que implica tareas como la estandarización de los datos, la generación de nuevas variables y la elección de características significativas. El éxito de las fases posteriores depende en gran medida de la calidad del trabajo realizado en esta etapa (Han, Kamber, & Pei, 2011).

4.4.4. Modelado

Dentro de este punto, se aborda y se aplican algoritmos de la minería de datos o el aprendizaje automático, con el fin de desarrollar modelos capaces de solucionar los problemas planteados. Dependiendo de los datos y los objetivos, se pueden implementar distintas técnicas, como son las redes neuronales, los árboles de decisión o modelos de regresión. Se puede llegar a probar varios modelos para comparar su rendimiento y determinar cuál es el más adecuado. En otras palabras, tiene como objetivo medir el rendimiento de los algoritmos empleados y los patrones o grupos encontrados en la fase de modelado (Tenés, 2023).

4.4.5 Evaluación

Una vez que se ha entrenado un modelo, es necesario evaluarlo para comprobar si cumple los objetivos establecidos. Esta fase incluye la revisión de métricas como la precisión y la exactitud del modelo, así como una validación cruzada para verificar que el modelo generalice bien a datos no vistos. Si el modelo no es satisfactorio, se pueden hacer ajustes y volver a fases anteriores para optimizarlo (Müller & Guido, 2016).

4.4.6 Despliegue

Esta fase de implementación se dispone a implementar el modelo en producción. Además, esta fase también suele incluir el monitoreo y mantenimiento del modelo a largo plazo. Y una característica fundamental de CRISP-DM es su flexibilidad y naturaleza iterativa. Las fases del proceso no son lineales, lo que permite volver a etapas anteriores para realizar ajustes cuando sea necesario. Esto es especialmente útil en proyectos donde los resultados iniciales requieren refinamiento a medida que se avanza. Esta adaptabilidad es una de las razones por las que CRISP-DM sigue siendo ampliamente utilizado en la industria (Heymann, (2021).

4.5 Metodología Scrum

Scrum es un proceso que propone la aplicación de una metodología de trabajo más colaborativa entre las áreas involucradas en el desarrollo de nuevos productos, con la finalidad de trabajar de manera holística y así lograr mejor comunicación, mayor integración y conocimiento por

parte de todos los roles para así hacer del trabajo más ágil, con productos efectivos y mayor satisfacción de los clientes (Salazar & Beltrán, C. A., 2020). Además, se plantea los tres ejes de Scrum; primeramente tenemos la Transparencia; misma que se refiere a los aspectos dentro del proceso, seguido de la Inspección debe ser revisada de manera constante según el manual de proceso con la finalidad de que la planeación sea eficiente y en tercer lugar la Revisión debe velar por cumplir todos los requisitos de aceptación (Garzon, Torres, & Vásquez, 2024).

4.6 Fundamentos tecnológicos

La integración de un sistema para la conversión de texto a voz (TTS) moderno implica el uso de diversas tecnologías y herramientas que han evolucionado significativamente en los últimos años. A continuación, se presentarán algunos de los más importantes para la realización de este proyecto.

4.6.1 Redes Neuronales Recurrentes (RNN) y Transformadores

Las redes neuronales recurrentes (RNN) han sido ampliamente utilizadas en la síntesis de voz debido a su capacidad para procesar secuencias temporales de datos (Goodfellow et al., 2016). Sin embargo, con la aparición de los transformadores, se ha observado un cambio hacia el uso de estos modelos en la generación de secuencias, incluidos los sistemas TTS, gracias a su mayor eficiencia y capacidad para captar dependencias de largo alcance en el contexto del lenguaje y el audio (Vaswani et al., 2017). Modelos como Whisper y SpeechT5, presentados en los últimos años, han implementado arquitecturas de transformadores para mejorar la fluidez y naturalidad en la generación de voz (Li et al., 2023).

Para este proyecto, se planea explorar ambas tecnologías, ya que es importante evaluar cuál de ellas proporciona una síntesis de voz más natural y adaptada a las necesidades del asistente robótico Volk, en especial en su capacidad para generar una prosodia empática. Se espera que los avances en los modelos basados en transformadores puedan ser aplicados en etapas posteriores del proyecto.

4.6.2 Frecuencias agradables y empatía en la voz sintética

Existen investigaciones recientes que señalan que la frecuencia de la voz puede influir en la percepción de emociones como empatía y simpatía. Además, un estudio llevado a cabo en 2023 por Lee et al. sugiere que las voces con tonos medios y frecuencias entre 180 Hz y 250 Hz tienden a ser percibidas como más cálidas y agradables para el oído humano, lo que facilita una conexión emocional más profunda (Lee, Kim, & Park, 2023). Además, estas observaciones son especialmente relevantes para sistemas como el asistente robótico Volk, que necesita generar una interacción que sea percibida como cercana y empática.

En este contexto, ajustar la frecuencia de la voz sintetizada a niveles que sean considerados más agradables puede ser una táctica fundamental para optimizar la aceptación y la satisfacción del usuario. De hecho, se ha encontrado que las variaciones en la prosodia, incluyendo el control de la frecuencia, la entonación y el ritmo, influyen significativamente en la percepción de la empatía en voces sintéticas (Smith et al., 2024).

4.6.3 Procesamiento del Lenguaje Natural (PLN)

El PLN o también nombrado como Procesamiento del Lenguaje Natural, es otra pieza fundamental en los sistemas TTS. Permite la transformación de entradas textuales en representaciones adecuadas para la síntesis de voz, garantizando que la salida no solo sea coherente en términos lingüísticos, sino que también respete las estructuras gramaticales del idioma. En el caso del proyecto Volk, será necesario adaptar los modelos de PLN al español de Latinoamérica para que puedan analizar y convertir el texto de manera eficiente, considerando las particularidades gramaticales y dialectales (Jurafsky & Martin, 2021). Esta adaptación se hará en fases posteriores, cuando el sistema de síntesis de voz esté más avanzado y se tenga mayor claridad sobre las necesidades específicas del usuario.

La síntesis de voz ha evolucionado significativamente en las últimas décadas, con el desarrollo de modelos avanzados que utilizan técnicas de aprendizaje profundo. Entre los modelos más destacados se encuentran Tacotron y WaveNet, que han establecido nuevos estándares en la calidad del audio generado (Hernández et al., 2021).

Tacotron y Tacotron 2 son arquitecturas que convierten texto en espectrogramas, que luego son transformados en audio. Estos modelos utilizan redes neuronales recurrentes (RNN) y permiten generar voces más naturales al capturar la entonación y la prosodia del habla humana (López et al., 2023).

Por otro lado, WaveNet, desarrollado por DeepMind, es un modelo generativo que crea formas de onda de audio a partir de espectrogramas, lo que mejora notablemente la calidad del sonido producido. Su arquitectura se basa en redes neuronales convolucionales que permiten modelar las dependencias temporales en el audio, logrando así una reproducción más fiel de la voz humana (Martínez et al., 2022).

La calidad del audio generado no solo depende del modelo elegido, sino también de las características prosódicas del dataset utilizado. En el contexto de la síntesis de voz, es esencial prestar atención a la variedad de las grabaciones y a la inclusión de expresiones emocionales que contribuyan a una mejor interacción humano-máquina (Pérez et al., 2023).

4.7 Perspectivas futuras y tecnologías emergentes en la síntesis de voz

En los últimos años, la investigación en la síntesis de voz ha avanzado hacia la creación de voces que no solo son naturalistas, sino también empáticas y adaptativas. Tecnologías emergentes como la combinación de sistemas TTS con modelos de predicción de emociones han comenzado a mostrar resultados prometedores. Por ejemplo, se han realizado estudios sobre la integración de modelos de voz que pueden cambiar su tono y prosodia en tiempo real, basados en el estado emocional detectado en el usuario (Kumar & Patel, 2024). Esta evolución sugiere que la empatía en los sistemas TTS no solo se logrará a través de la prosodia preentrenada, sino también mediante la capacidad del sistema para ajustarse dinámicamente a las interacciones humanas.

V. Marco metodológico.

Para el desarrollo del sistema de conversión de texto a voz, se implementaron algunas estrategias que nos permitieron estructurar nuestro proyecto de una manera mucho más organizada y flexible. A lo largo del proyecto, se aplicaron metodologías que facilitaron el análisis de datos, la optimización del modelo y la gestión eficiente del desarrollo.

Dado que en el Objetivo 2 se abordó la importancia de estas estrategias en la planificación y ejecución del sistema, a continuación, se presentan en detalle las metodologías aplicadas, explicando su implementación y su impacto en cada fase del desarrollo.

5.1 Metodologías aplicadas

Primer enfoque CRISP-DM

Para estructurar el desarrollo del sistema de conversión de texto a voz, se utilizó CRISP-DM, un modelo de proceso ampliamente aplicado en proyectos de minería de datos (Schröer, Kruse y Gómez, 2021). Esta metodología permitió trabajar de forma iterativa y flexible, asegurando ajustes constantes tanto en el modelo como en los datos a lo largo de las distintas etapas del proyecto. Gracias a esto, se logró una mejora continua, especialmente en lo relacionado con la prosodia empática, aspecto clave para hacer la voz sintetizada más natural y expresiva. Además, se considera importante mencionar los pasos desarrollados.

Entendimiento del negocio: Desde el inicio, el objetivo fue claro: el proyecto debía generar una voz en español de Latinoamérica que fuera natural, fluida y empática. Para lograrlo, fue necesario analizar qué características debía tener la voz para mejorar la interacción humano-máquina. En tal virtud, se definieron aspectos esenciales como la entonación, el ritmo y la expresividad, asegurando que el habla sintetizada no sólo fuera comprensible, sino también agradable de escuchar y capaz de transmitir emociones.

Análisis de datos: Con el objetivo bien establecido, el siguiente paso fue estudiar las particularidades del español latinoamericano. Se analizaron elementos como la entonación, el acento y la variabilidad en la pronunciación.

Preparación de los datos: Una vez de haber analizado, se creó un dataset propio, compuesto por grabaciones de frases pronunciadas por hablantes nativos. Para garantizar la calidad del entrenamiento, estos audios fueron preprocesados, eliminando ruidos de fondo, normalizando el volumen y asegurando una dicción clara y uniforme. Esta fase fue crucial, ya que la calidad del dataset impacta directamente en la precisión y naturalidad del modelo final.

- **Modelado:** Se diseñó un modelo de aprendizaje profundo que transformó el texto en audio, ajustando la prosodia para transmitir emociones y empatía efectivamente.
- **Evaluación:** La calidad de la voz sintetizada se midió mediante pruebas con usuarios, con un enfoque especial en la percepción de la naturalidad y empatía de la voz.
- **Despliegue:** El modelo final fue preparado con los estándares necesarios para su eventual integración en el asistente robótico Volk, garantizando su compatibilidad con entornos reales.

Segundo enfoque scrum

Scrum es una metodología ampliamente utilizada en el desarrollo de software, permitiendo una gestión estructurada y adaptable de los proyectos (Morandini, Oliveira Jr, Corrêa, 2021). En este caso, facilitó el desarrollo del sistema de conversión de texto a voz mediante ciclos cortos de trabajo, asegurando que cada fase del proyecto se ajustara de manera iterativa y progresiva.

A continuación, se detallan los aspectos clave de su implementación en el proyecto.

Planificación de sprints

El trabajo se organizó en sprints de dos semanas, con entregables funcionales en cada fase:

1. **Creación del dataset:** Preprocesamiento y organización de muestras de audio.
2. **Entrenamiento inicial:** Generación de voz con la red neuronal.
3. **Ajustes en la prosodia y empatía:** Optimización de la entonación y expresividad.
4. **Validación del sistema:** Pruebas finales con distintos textos.

Se realizaron **revisiones semanales** para evaluar avances y corregir errores en el entrenamiento del modelo. Dentro del equipo:

- **Product Owners:** Definieron objetivos y prioridades (Kaar Joseph Mashingashi Unkuch y Juan Sebastián Andrade Alulema).
- **Scrum Master:** Facilitó la metodología y eliminó obstáculos (Ing. Cristian Timbi).
- **Equipo de Desarrollo:** Implementó, probó y ajustó el sistema (Kaar Joseph Mashingashi Unkuch y Juan Sebastián Andrade Alulema).

El desarrollo del proyecto combinó CRISP-DM y Scrum, asegurando un enfoque estructurado y flexible en la creación del módulo TTS. Para validar la calidad del modelo, se diseñó y aplicó una encuesta dirigida a expertos en el área, con el objetivo de evaluar la percepción sobre la voz generada en un asistente robótico socialmente interactivo.

La encuesta fue aplicada a Ordoñez Vásquez María Elisa (Anexo 2.1), Reyes Quezada Karina Priscila (Anexo 2.2), Almeida Soliz Nidia Elizabeth (Anexo 2.3) y Zamora Torres Johana Elizabeth (Anexo 2.4).

Este análisis combinó datos cuantitativos y cualitativos, permitiendo identificar áreas de mejora y optimizar la voz sintetizada para su implementación en el asistente robótico.

Scrum permitió una organización eficiente del desarrollo, asegurando mejoras iterativas y validaciones constantes. La retroalimentación de expertos ayudó a optimizar la expresividad del modelo y garantizar su calidad para la implementación en el asistente robótico.

5.2 Evaluación de modelos de aprendizaje profundo para TTS (OE1)

Para seleccionar el modelo de conversión de texto a voz, se evaluaron varias arquitecturas en base a la entonación, a la fluidez y la claridad en la pronunciación. También se probaron

modelos como Tacotron 2, CoquiTTS, MozillaTTS, Bark y Google Text-to-Speech para analizar cuál modelo era el que lograba una síntesis más natural, mucho más estable y mejor.

Después de varias pruebas y de varios análisis detallados, los modelos que se descartaron fueron Bark y Google Text-to-Speech. Bark, aunque estos lograban tener una síntesis aceptable, en algunos y ciertos casos, se presentaban algunos pequeños problemas de consistencia en la entonación, entonces debido a esto, los modelos generaban variaciones poco naturales en el habla. Por otro lado, Google Text-to-Speech, si bien llegaba a producir audios claros y bien articulados, tenía una prosodia rígida y poco expresiva, lo que hacía que la voz sonara más mecánica y menos adaptable al contexto.

Lo cual al final de todos estos análisis a fondo, nos centramos 3 modelos específicos, el cual son **Tacotron 2, MozillaTTS y CoquiTTS**, estos modelos mostraron mejor desempeño en términos de naturalidad y estabilidad en la modulación de la voz. Entonces a partir de estos tres, se realizó una evaluación mucho más a fondo y detallada, para así determinar cuál modelo es el que ofrecía la mejor síntesis de voz.

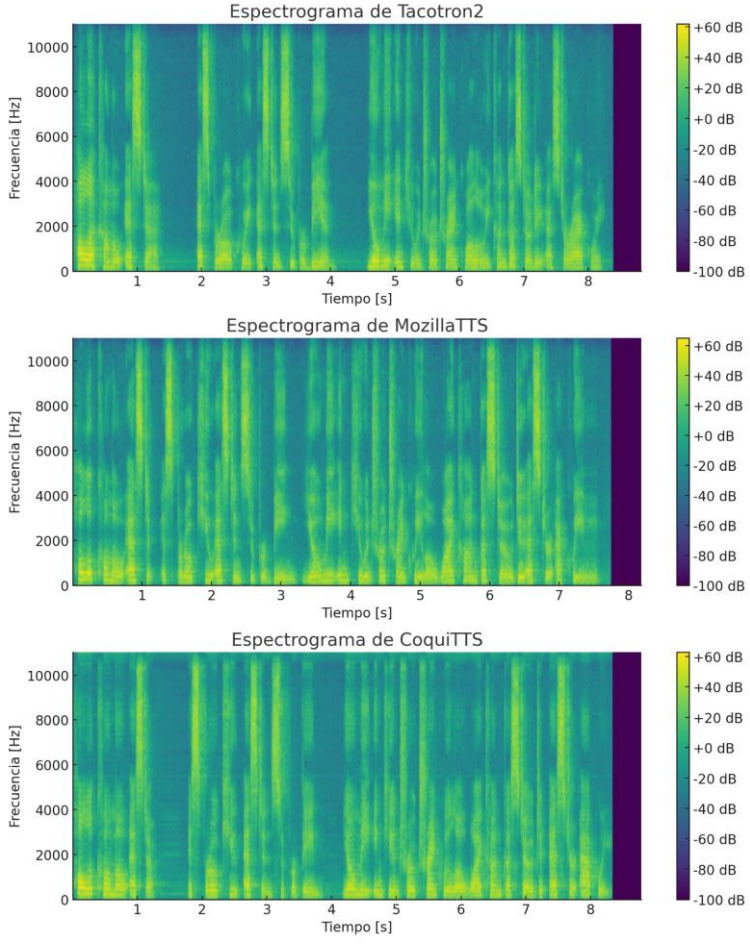


Ilustración 1: *Espectrogramas para la elección de Tacotron 2*

Al momento de analizar estos espectrogramas que se muestra en la **Ilustración 1**, se identificaron diferencias clave entre los espectrogramas de Tacotron 2, MozillaTTS y

CoquiTTS, lo que permitió analizar la selección del modelo más adecuado para la síntesis de voz.

Para el espectrograma de **Tacotron 2**, se observó una distribución de sus frecuencias más uniforme, con varias transiciones suaves las cuales se encuentran bien definidas entre los fonemas. Sus bandas de energía que son las frecuencias altas y medias se mantuvieron consistentes a lo largo del tiempo, reflejando una buena pronunciación, clara y continua. Esto indicó que el modelo generaba un habla más fluida, sin cortes abruptos ni irregularidades en la modulación.

Pero para el modelo siguiente, en el espectrograma de **MozillaTTS,** se llegó a notar variaciones irregulares en la intensidad de las frecuencias, lo que evidenció una entonación inestable. Además, las transiciones entre segmentos de voz resultaron menos definidas, lo que generaba una percepción de habla menos natural o incluso robótica.

Por último tenemos, el espectrograma de **CoquiTTS**, este presenta varios patrones de energía con cortes mucho más marcados y pausas abruptas. Estos patrones nos indicaron que el modelo no logró tener una buena continuidad y no tenía una fluidez en la entonación, lo que provocó que la voz generada sonaba entrecortada o con pausas inesperadas.

Entonces, gracias a este análisis, el modelo de Tacotron 2 nos mostró una mejor segmentación de fonemas y una entonación más estable, así llegando a evitar los problemas observados en MozillaTTS y CoquiTTS. Su fluidez en la modulación y su uniformidad en la distribución de frecuencias evidenciaron que este modelo produjo una síntesis de voz más natural, convirtiéndo este modelo en la mejor opción para este proyecto.

5.3 Desarrollo del módulo de transcripción y conversión de texto a voz (OE2)

En esta sección del proyecto, y su implementación del sistema, se estructuró su desarrollo en varias capas la cual optimizaban el procesamiento y así garantizaban una conversación fluida de texto a voz.

Para la **capa de presentación**, se realizó una interfaz que permitió cargar archivos .wav, una vez cargado, permitió visualizar transcripciones y reproducir el audio generado. En la **capa de entrada y preprocesamiento**, se integró Whisper para la transcripción automática y se aplicaron técnicas de eliminación de ruido y normalización, mejorando la calidad del texto extraído.

Para la **capa de procesamiento**, se configuró el Tacotron 2 para asi transformar texto en espectrogramas de Mel, pudiendo ajustar hiperparámetros clave para asi poder llegar a optimizar la precisión del modelo. Continuamos con la **capa de síntesis de voz**, en esta capa se utilizaron estos espectrogramas para así generar audio realista con HiFi-GAN, aplicando filtros adicionales para mejorar la calidad del sonido.

En todo este proceso, se pudo cumplir con el Objetivo Específico 2 (OE2), pudiendo lograr una síntesis de voz clara y natural adecuada para el proyecto, facilitando así su integración y la generación de diálogos más expresivos, como se muestra en la **Ilustración 2.**

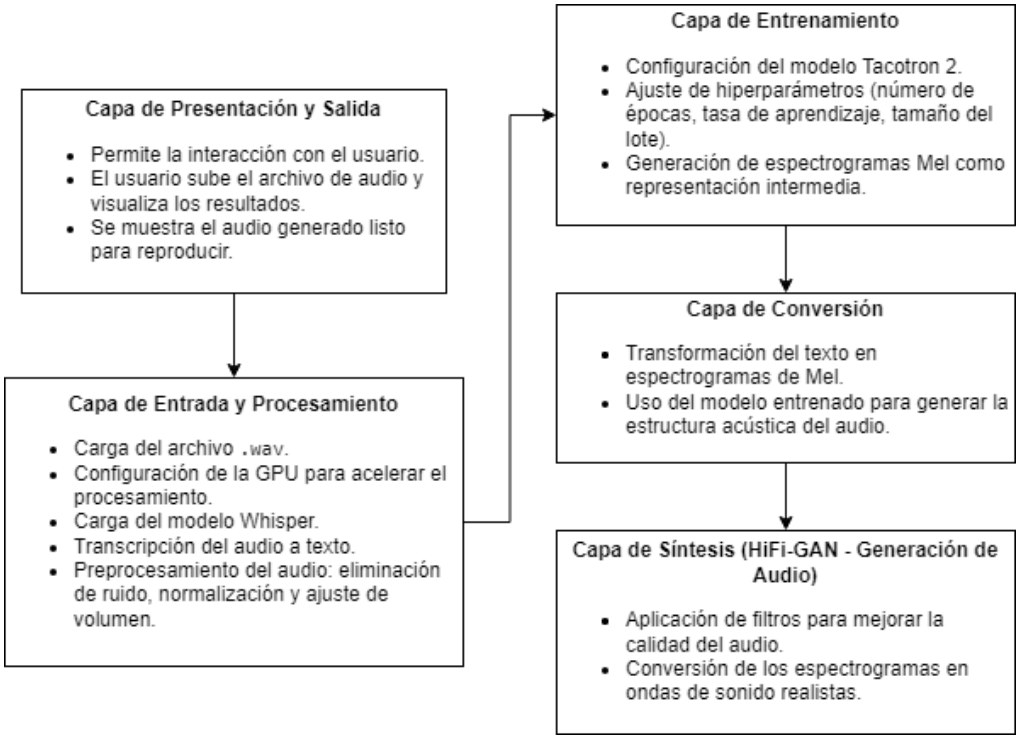


Ilustración 2: *Diagrama de flujo para Transcripción y conversión TTS.*

De igual manera se definió una arquitectura estructurada, la cual está representada en la **Ilustración 3.** Esta organización nos permitió que cada módulo contribuyera al cumplimiento de los objetivos específicos.

Dentro de la **capa de presentación**, como ya se mencionó previamente, se diseñó una interfaz que facilitó la carga de archivos .wav, la visualización de transcripciones y la reproducción del audio sintetizado, teniendo así una gran interacción intuitiva y mucho más accesible con los usuarios cumpliendo así el OE3.

En la **capa de entrada y procesamiento,** se integró Whisper para la transcripción automática de audio a texto, este está acompañado de técnicas de preprocesamiento como lo son la eliminación de ruido y la normalización de frecuencia. Además de esto, se llegó a implementar dependencias y herramientas como lo son Python, Librosa y Torch para mejorar la calidad del procesamiento cumpliendo así con el OE1.

Para la **capa de entrenamiento**, el Tacotron 2 que se configuró para generar espectrogramas de Mel, se fue ajustando con hiperparámetros, para así optimizar la representación fonética del audio. Se almacenó el modelo entrenado para su reutilización y eficiencia futura teniendo en cuenta el OE2

En la **capa de conversión**, Tacotron 2 transformó los textos en espectrogramas de Mel, con estos, asegurando una conversión precisa de la estructura lingüística a una representación acústica óptima cumpliendo y tomando en cuenta el OE2.

Por último, en la **capa de síntesis**, HiFi-GAN convirtió los espectrogramas en ondas sonoras realistas, aplicando filtros para mejorar la calidad del sonido y generando archivos listos para su evaluación. Esta etapa permitió validar la naturalidad e inteligibilidad del audio generado y teniendo en cuenta el OE3.

Como se observa en la **Ilustración 3**, esta arquitectura optimizó la conversión de texto a voz, asegurando una síntesis clara y precisa.

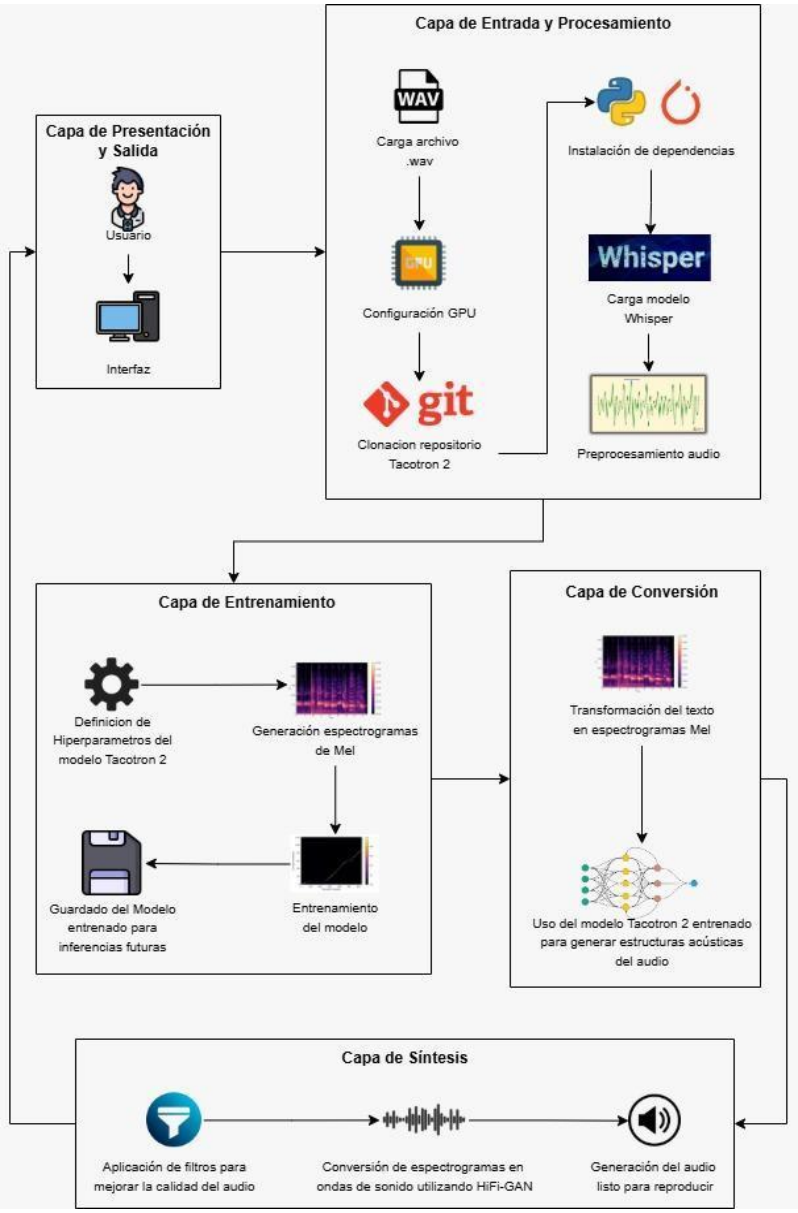


Ilustración 3: *Diagrama de Arquitectura de software.*

Dadas las ilustraciones anteriores, es posible entender con mayor claridad los elementos clave de esta sección. Ahora, avanzaremos con un desarrollo detallado, explicando cada aspecto de forma clara y estructurada para asegurar una comprensión completa.

El desarrollo del sistema de conversión de texto a voz (TTS) se llevó a cabo mediante un enfoque modular y estructurado, asegurando que cada fase se integrará de manera eficiente. Para ello, se inició con la configuración del entorno, se instalaron las bibliotecas y herramientas necesarias para garantizar la compatibilidad con Tacotron 2 y el vocoder HiFi-GAN, optimizando así el procesamiento de datos y la ejecución del modelo.

Posteriormente, se recopilaron y se preprocesaron los datos, se construyó un dataset propio adaptado a los requerimientos del asistente robótico Volk. Se aplicaron técnicas de limpieza, eliminación de ruido, normalización de volumen y conversión de las muestras en espectrogramas Mel, asegurando la calidad y coherencia de los datos de entrada para el entrenamiento del modelo.

El desarrollo del modelo se basó en la arquitectura de Tacotron 2, se ajustaron hiperparámetros como la tasa de aprendizaje, el tamaño del lote y el número de épocas, con el objetivo de mejorar la estabilidad y la prosodia del habla sintetizada. El entrenamiento fue iterativo, se evaluaron los espectrogramas generados para refinar la claridad y expresividad del audio.

Para garantizar un manejo eficiente de los datos y el desarrollo del modelo, se adoptó **la metodología CRISP-DM,** lo que permitió estructurar cada etapa del proceso, desde la adquisición de datos hasta su integración en el sistema de síntesis de voz. Este enfoque aseguró que el modelo operara con información bien estructurada y optimizada.

Con la metodología definida y el modelo implementado, se realizó un proceso exhaustivo de gestión de datos; se alineó la información utilizada en el entrenamiento con los objetivos del asistente robótico Volk. Se establecieron criterios específicos para garantizar que la voz sintetizada fuera clara, natural y empática, priorizando la entonación y la prosodia para lograr una comunicación efectiva en entornos educativos y de asistencia personalizada.

A partir de este punto, el manejo de datos se abordó en distintas fases, asegurando que cada etapa contribuyera a la mejora de la síntesis de voz.

- **Entendimiento del negocio:**
    - Se definieron y documentaron los objetivos del asistente robótico Volk, enfocándose en el desarrollo de un sistema de conversión de texto a voz (TTS) en español latinoamericano. Se priorizó que la voz generada fuera clara, natural y empática, adaptándose a las necesidades comunicativas de los usuarios.
    - Se establecieron los requerimientos clave para que el módulo pudiera funcionar en entornos educativos y de asistencia personalizada, asegurando que la entonación y la prosodia del habla fueran lo más naturales posible.

- **Análisis de datos:**
    - Se analizaron las propiedades prosódicas y dialectales del español latinoamericano con el objetivo de identificar patrones clave que garantizaran una voz sintetizada natural y comprensible. Se evaluaron aspectos como la entonación, el ritmo y la variabilidad fonética, fundamentales para mejorar la calidad de la conversión de texto a voz.
    - Para el procesamiento y análisis de los datos, se utilizaron herramientas especializadas como Librosa, que permitió la manipulación y extracción de características de los archivos de audio, y Matplotlib, empleada para la visualización y análisis de los espectrogramas y frecuencias de sonido. Estas herramientas facilitaron el estudio detallado de los datos acústicos y su adaptación al modelo de síntesis de voz.

- **Preparación de los datos:**
    - Se desarrolló un dataset propio conformado inicialmente por 100 grabaciones de audio, cada una con una duración de 9 a 15 segundos en español latinoamericano. Estas grabaciones fueron seleccionadas y diseñadas estratégicamente para representar diversas entonaciones y emociones, con el fin de garantizar una mayor naturalidad y expresividad en la síntesis de voz.
    - El proceso de recopilación y estructuración de los datos fue clave para proporcionar al modelo un conjunto de entrenamiento robusto, capaz de capturar las variaciones fonéticas y prosódicas necesarias para mejorar la calidad de la conversión de texto a voz.

- **Preprocesamiento:** Para el análisis del preprocesamiento tenemos cuatro pasos a seguir; mismo que detallamos a continuación:
    - **Eliminación de ruido:** Se implementaron filtros digitales avanzados con el objetivo de reducir interferencias y mejorar la claridad del audio, asegurando que la señal de voz se mantuviera limpia y comprensible.
    - **Normalización del volumen:** Se ajustaron los niveles de volumen de cada grabación para mantener la coherencia acústica dentro del dataset, evitando fluctuaciones abruptas que pudieran afectar la estabilidad del entrenamiento.
    - **Conversión a espectrogramas Mel:** Se transformaron los archivos de audio en representaciones espectrales Mel, que constituyeron la entrada primaria del modelo Tacotron 2, facilitando la extracción de patrones fonéticos y la generación de voz sintética más natural.
    - **Estandarización de nombres:** Los archivos de audio fueron renombrados con identificadores numéricos, permitiendo una mejor organización del dataset y facilitando su procesamiento automatizado en las siguientes fases del desarrollo.

Además, para optimizar aún más la calidad del dataset, se llevó a cabo un **proceso de reducción de ruido** utilizando Audacity, asegurando que las señales de audio estuvieran libres de interferencias que pudieran afectar la precisión de la síntesis de voz. Se aplicaron tres parámetros fundamentales en este proceso.

- **Reducción de ruido (16 dB):** Se definió este valor para minimizar interferencias sin comprometer la inteligibilidad del habla.
- **Sensibilidad (8.0):** Se estableció un umbral que equilibrara la eliminación del ruido de fondo sin afectar detalles esenciales del habla.
- **Suavizado de frecuencia (12 bandas):** Se empleó para distribuir uniformemente las modificaciones en las frecuencias, evitando alteraciones abruptas en el espectro de audio.

La aplicación de estas técnicas garantizó que el dataset estuviera limpio y optimizado, lo que permitió que el modelo Tacotron 2 aprendiera de manera más eficiente las características fonéticas y prosódicas del español latinoamericano. Como se muestra en las ilustraciones 4, 5 y 6, la implementación de este procedimiento mejoró la claridad y naturalidad de la señal de voz, contribuyendo a un mejor desempeño del modelo durante la fase de entrenamiento.

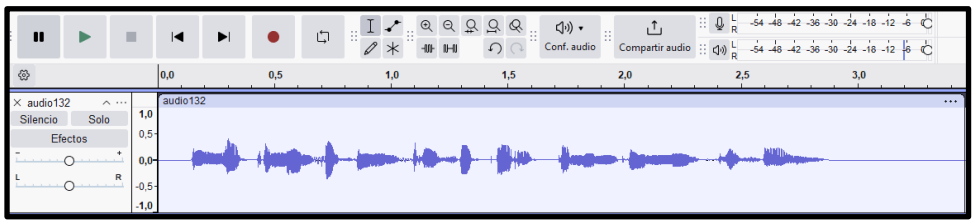


Ilustración 4: *Audio sin procesar en Audacity*

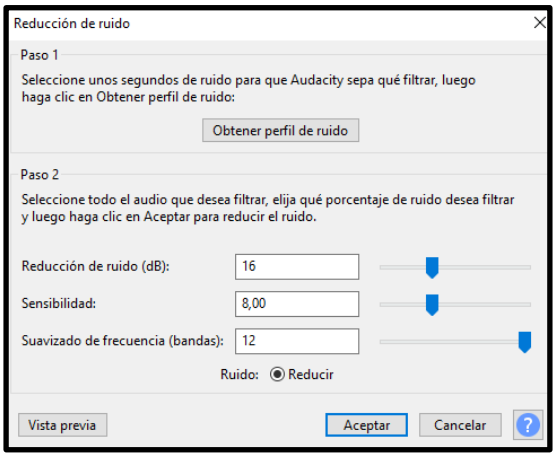


Ilustración 5: *Configuración para eliminación de ruido*

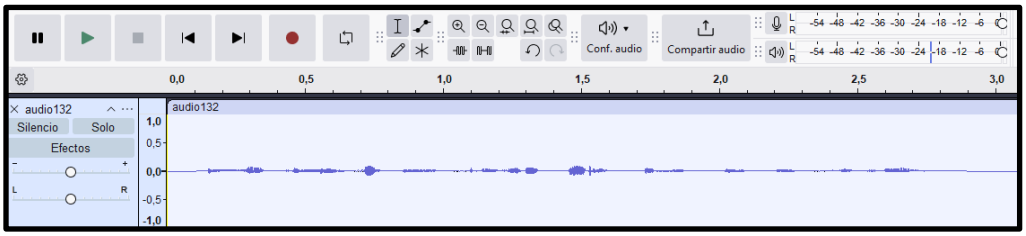


Ilustración 6: *Resultado del audio eliminando ruido*

Con los archivos de audio debidamente limpiados y optimizados, el siguiente paso fue su exportación en formato WAV con una frecuencia de muestreo de 22050 Hz, como se mostró en la Ilustración 6. Esta configuración fue seleccionada debido a su compatibilidad con el modelo de síntesis de voz y su capacidad para preservar la información acústica sin degradación perceptible. Mantener una frecuencia de muestreo estándar permitió garantizar una reproducción fiel de las características fonéticas y prosódicas, asegurando que la síntesis de voz reflejara con precisión las variaciones naturales del habla.

Además, se estableció que los archivos de audio fueran almacenados en modo mono, en lugar de estéreo. Esta decisión respondió a la necesidad de optimizar la claridad del sonido y reducir la carga computacional, evitando información redundante en el procesamiento del modelo. Al utilizar un solo canal, se logró disminuir el tamaño de los archivos sin afectar la inteligibilidad del habla, lo que resultó beneficioso tanto para la eficiencia del sistema como para la gestión de recursos durante la fase de entrenamiento.

Para complementar esta configuración, se definió un bitrate de 170-210 kbps, equilibrando la calidad del audio y el tamaño de los archivos, tal como se apreció en la Ilustración 6. Este ajuste fue fundamental para evitar que los archivos de gran tamaño afectaran el rendimiento del sistema, garantizando un flujo de datos ágil y eficiente. Minimizar la carga de almacenamiento y procesamiento permitió que el modelo de aprendizaje profundo manejara la información sin ralentizar la ejecución, optimizando así el proceso de entrenamiento.

Gracias a estas configuraciones, el dataset quedó estructurado de manera óptima para alimentar el modelo de conversión de texto a voz (TTS). La combinación de reducción de ruido, estandarización de la frecuencia de muestreo y simplificación del formato de audio permitió generar datos de entrada más precisos, asegurando un aprendizaje más eficiente. Estas optimizaciones no solo contribuyeron a la estabilidad del modelo, sino que también tuvieron un impacto directo en la naturalidad, fluidez e inteligibilidad del audio generado, mejorando significativamente la calidad de la síntesis de voz. Ver ilustración 7.

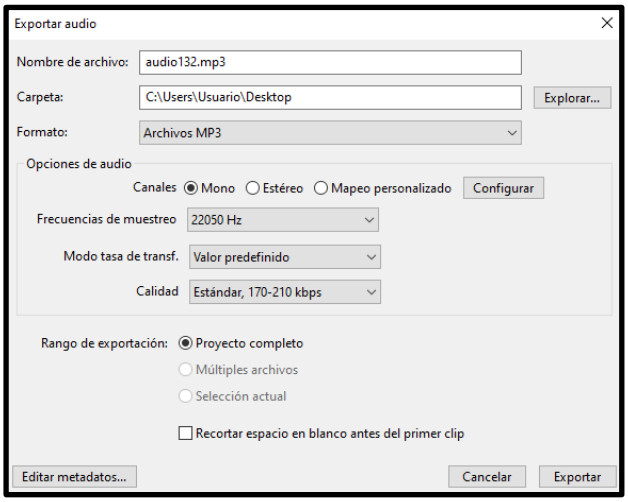


Ilustración 7: *Configuraciones para guardar el audio procesado*

Con el conjunto de datos optimizado y estructurado, el siguiente paso fue la **implementación y entrenamiento del modelo de conversión de texto a voz.** Para ello, se seleccionó Tacotron 2, un modelo basado en aprendizaje profundo ampliamente utilizado en la síntesis de voz. Su elección se debió a su capacidad para generar una entonación natural, una prosodia fluida y una alta inteligibilidad, superando las limitaciones de modelos previos como Tacotron 1 y WaveNet, que presentaban desafíos en la alineación y estabilidad del habla generada.

El modelo se diseñó con dos módulos fundamentales que trabajaron en conjunto para transformar el texto en una representación acústica, estos módulos son los siguientes:

- **Codificador (Encoder):** Extrajo características fonéticas y contextuales del texto de entrada, convirtiendo caracteres y fonemas en una representación numérica que pudiera ser procesada por el modelo.
- **Decodificador (Decoder):** Generó los espectrogramas de Mel a partir de la información obtenida en el codificador. Para ello:
    - Se implementó un mecanismo de atención sensible a la ubicación, que permitió asignar pesos a cada parte del texto, asegurando que el modelo se enfocara en la sección correcta en cada paso de la síntesis.
    - Se utilizó redes LSTM unidireccionales con 1024 neuronas, encargadas de procesar la información secuencialmente y generar la salida acústica en forma de espectrogramas.
    - Se incorporó una capa de predicción de finalización (Gate Layer), diseñada para indicar cuándo debía detenerse la generación del audio, evitando la producción de sonidos innecesarios o prolongaciones artificiales.

Una vez generados los espectrogramas de Mel, se requirió un vocoder para convertirlos en una forma de onda realista. En este caso, se empleó HiFi-GAN, un modelo de redes neuronales convolucionales capaz de sintetizar audio con un nivel de realismo superior a los métodos tradicionales como Griffin-Lim. Su estructura, basada en 13 capas convolucionales, permitió

que la conversión de los espectrogramas a señales de audio se ejecutaron con alta calidad y estabilidad.

Para garantizar un entrenamiento eficiente, se definieron los siguientes parámetros iniciales:

- **Tasa de aprendizaje: 0.001**: Se estableció este valor para equilibrar la velocidad de entrenamiento y la estabilidad del modelo. Un valor más alto podría haber generado inestabilidad, mientras que una tasa demasiado baja habría ralentizado la convergencia.
- **Tamaño del lote: 32**: Se seleccionó este tamaño para optimizar el rendimiento computacional sin comprometer la estabilidad del modelo. Los lotes más grandes habrían requerido más memoria, mientras que lotes más pequeños podrían haber afectado la convergencia del aprendizaje.
- **Número de épocas: 250:** A partir de pruebas experimentales, se determinó que después de 200 épocas, la calidad de la síntesis mejoraba notablemente, sin riesgo de sobreajuste.

Durante el entrenamiento, se analizaron constantemente los espectrogramas de Mel generados en cada iteración para evaluar la calidad del modelo y detectar posibles desajustes en la síntesis de voz. En la Ilustración 7, se muestra un espectrograma obtenido en las primeras etapas del entrenamiento, el cual refleja la evolución del modelo en la generación del habla.

El análisis de estos espectrogramas permitió extraer información clave sobre la síntesis de voz:

- **Las zonas más brillantes indicaron** regiones con mayor energía en la señal de audio, generalmente asociadas a fonemas de alta intensidad.
- **Las líneas verticales representan** la estructura temporal de los fonemas, evidenciando la segmentación del habla sintetizada.
- **Las áreas oscuras reflejaron** pausas y momentos de menor energía, fundamentales para la fluidez y naturalidad del discurso.

A partir de estos análisis, se identificaron algunos desajustes en la transición entre fonemas y en la entonación de frases largas, lo que llevó a realizar ajustes adicionales en la configuración del modelo. Se optimizaron las capas convolucionales y el mecanismo de atención, logrando una mayor fluidez en el habla y una mejor alineación entre el texto y el audio.

Este proceso de refinamiento permitió mejorar significativamente la calidad de la síntesis de voz, asegurando que la entonación, la segmentación y la expresividad del habla generada se asemejara lo más posible a una voz natural y humanizada como se observa en la ilustración 8.

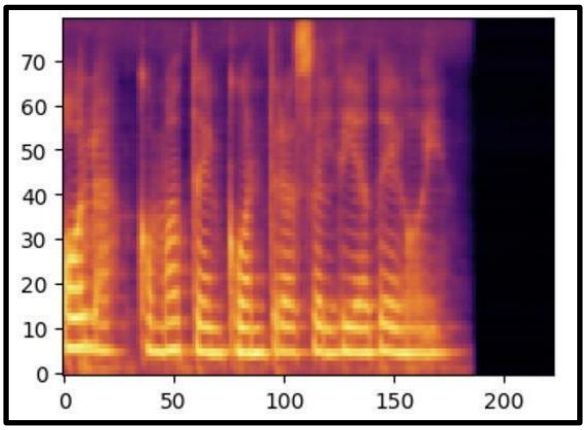


Ilustración 8: *Espectrograma de Mel de resultados preliminares*

Para garantizar una implementación eficiente y optimizada del sistema de conversión de texto a voz, se adoptó un enfoque basado en **sprints iterativos bajo la metodología ágil Scrum.** Este enfoque permitió estructurar el desarrollo en fases manejables, asegurando una integración progresiva y reduciendo los riesgos asociados a errores acumulativos.

Cada sprint se diseñó con objetivos específicos alineados con la arquitectura del modelo y la infraestructura del sistema, permitiendo evaluar constantemente el desempeño del modelo Tacotron 2. Gracias a esta metodología, se logró una adaptación dinámica de los componentes del sistema, asegurando que cada iteración aportara mejoras tangibles a la calidad de la síntesis de voz.

Durante la ejecución de los sprints, se llevaron a cabo diversas actividades clave para la optimización del modelo:

- **Definición y planificación de objetivos específicos:** Cada iteración se inició con una planificación detallada, asegurando que los ajustes estuvieran alineados con la evolución del modelo y las métricas de calidad establecidas.
- **Implementación y optimización de módulos funcionales:** Se realizaron ajustes en los parámetros del sistema y se integraron nuevos componentes para mejorar la entonación, fluidez y estabilidad de la síntesis de voz.
- **Evaluación del rendimiento del modelo:** Se llevaron a cabo pruebas comparativas basadas en el análisis de espectrogramas de Mel, validando la inteligibilidad y naturalidad del audio sintetizado.
- **Documentación de resultados y ajustes en los parámetros:** Cada iteración fue documentada para mantener la trazabilidad del desarrollo, permitiendo una optimización basada en datos y asegurando mejoras progresivas en la calidad del sistema.

Entonces gracias a este enfoque iterativo, se logró perfeccionar la síntesis de voz paso a paso, asegurando que cada fase contribuye al refinamiento del modelo. A medida que avanzaron los sprints, se identificaron patrones de mejora en la alineación entre texto y audio, lo que llevó a

optimizaciones en la modulación de la entonación y la reducción de errores en la generación del habla.

Este proceso permitió no solo una implementación estructurada del modelo, sino también una evaluación constante de su desempeño, facilitando la toma de decisiones basada en resultados cuantificables. La metodología ágil aplicada en el desarrollo garantizó que cada iteración aportara mejoras significativas, optimizando la conversión de texto a voz y acercándose cada vez más a una síntesis natural y expresiva.

Como parte del desarrollo iterativo del sistema, se implementaron sprints que permitieron estructurar las actividades de manera organizada y optimizar cada fase del proceso. En el primer sprint, el enfoque estuvo en la configuración del entorno de trabajo, asegurando que la infraestructura estuviera correctamente preparada para el procesamiento de datos de audio y el entrenamiento del modelo de síntesis de voz.

Para garantizar la estabilidad del sistema y la compatibilidad con las herramientas necesarias, se estableció un entorno basado en Python 3.9.20, utilizando Anaconda como gestor de entornos virtuales. Esta elección permitió una **administración eficiente de las dependencias**, facilitando la integración de las librerías esenciales para el desarrollo del modelo Tacotron 2.

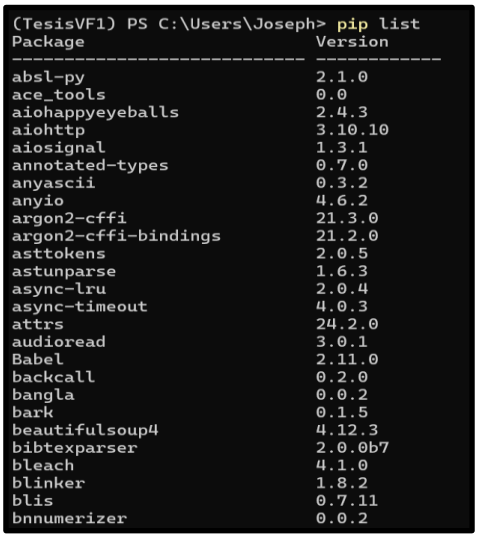
Durante esta fase, se instalaron y configuraron las siguientes librerías clave:

- **TensorFlow 2.18.0:** Fundamental para la implementación de redes neuronales profundas y la optimización del modelo Tacotron 2.
- **Librosa 0.9.2:** Utilizada para el procesamiento de señales de audio y la extracción de características acústicas esenciales para el entrenamiento del sistema.
- **Matplotlib 3.8.4:** Empleada en la visualización de espectrogramas de Mel y otros análisis gráficos que permitieron evaluar la calidad del audio sintetizado.
- **TTS 0.22.0:** Librería clave en la integración del modelo Tacotron 2, facilitando la conversión de texto a voz dentro del pipeline de procesamiento.
- **Torch 2.5.1:** Optimizada para la ejecución del modelo en GPU, mejorando significativamente los tiempos de procesamiento y entrenamiento.

Una vez completada la instalación de estas dependencias, se verificó su correcta integración mediante el comando *pip list*, asegurando que cada paquete estuviera correctamente instalado y compatible con el entorno de trabajo. Posteriormente, se realizaron pruebas preliminares con modelos preentrenados de síntesis de voz en español latinoamericano, con el objetivo de validar la infraestructura antes de proceder con la personalización y ajuste del modelo.

Esta primera fase fue esencial para garantizar la estabilidad del sistema y la disponibilidad de las herramientas necesarias para las siguientes etapas del desarrollo. La correcta configuración del entorno no solo facilitó la implementación del modelo, sino que también permitió optimizar el flujo de trabajo, asegurando que el proceso de entrenamiento y síntesis de voz se ejecutara de manera eficiente.

En la Ilustración 8, se presentó un resumen de las librerías utilizadas en esta configuración inicial, evidenciando la base tecnológica sobre la cual se construyó el sistema de conversión de texto a voz. Ver ilustración 9.



```
(TesisVF1) PS C:\Users\Joseph> pip list
Package                   Version
------------------------- -----------
absl-py                   2.1.0
ace_tools                 0.0
aiohappyeyeballs          2.4.3
aiohttp                   3.10.10
aiosignal                 1.3.1
annotated-types           0.7.0
anyascii                  0.3.2
anyio                     4.6.2
argon2-cffi               21.3.0
argon2-cffi-bindings      21.2.0
asttokens                 2.0.5
astunparse                1.6.3
async-lru                 2.0.4
async-timeout             4.0.3
attrs                     24.2.0
audioread                 3.0.1
Babel                     2.11.0
backcall                  0.2.0
bangla                    0.0.2
bark                      0.1.5
beautifulsoup4            4.12.3
bibtexparser              2.0.0b7
bleach                    4.1.0
blinker                   1.8.2
blis                      0.7.11
bnnumerizer               0.0.2
```

Ilustración 9: *Librerías necesarias instaladas*

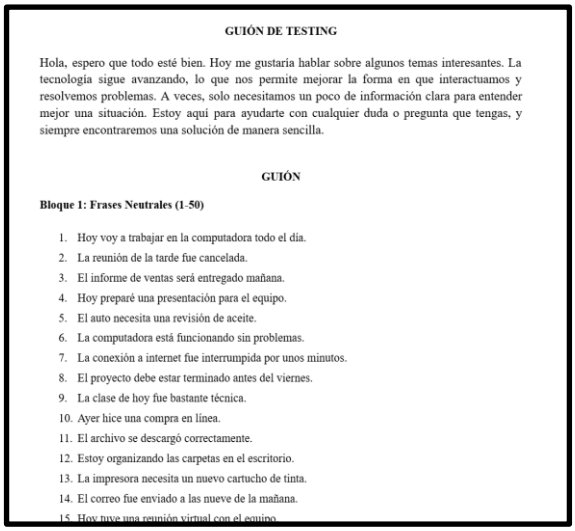
Tras completar la configuración del entorno y la integración de las herramientas necesarias, el siguiente paso en el desarrollo del sistema fue la creación del dataset y el preprocesamiento de los datos, un componente clave para garantizar la calidad del modelo de síntesis de voz. En esta fase, se diseñó un conjunto de datos personalizado, adaptado específicamente para la generación de voz en español latinoamericano, asegurando que el modelo pudiera aprender con ejemplos claros, expresivos y representativos de la lengua.

Para construir un dataset robusto y diverso, se elaboró un **guión lingüísticamente variado**, que incluyó una amplia selección de palabras, frases y expresiones que reflejaban distintos registros fonéticos, prosódicos y emocionales. Esta estrategia permitió capturar con precisión las variaciones naturales del español latinoamericano, como los cambios de entonación, ritmo y pausas, aspectos esenciales para lograr una síntesis de voz realista y fluida. En la Ilustración 10, se ilustró la estructura del guión utilizado, el cual sirvió como base para la grabación de 300 audios, garantizando una cobertura completa de las variaciones lingüísticas necesarias para el entrenamiento del modelo.

Más allá de la diversidad fonética, se priorizó la calidad del audio en cada grabación. Para ello, se establecieron criterios estrictos en la captura de sonido, asegurando que las voces fueran claras, sin ruido de fondo y con una dicción precisa. Estos factores fueron determinantes para que el modelo Tacotron 2 pudiera generar una voz más natural y comprensible en la etapa de inferencia.

La construcción de este dataset representó un paso fundamental en el desarrollo del sistema, ya que la calidad de los datos de entrenamiento impactó directamente en la fluidez, expresividad e inteligibilidad del habla sintetizada. Contar con un conjunto de datos bien estructurado

permitió mejorar la estabilidad del modelo, minimizando errores en la generación del habla y asegurando que la conversión de texto a voz fuera lo más realista posible. Ver ilustración 10.



GUIÓN DE TESTING

Hola, espero que todo esté bien. Hoy me gustaría hablar sobre algunos temas interesantes. La tecnología sigue avanzando, lo que nos permite mejorar la forma en que interactuamos y resolvemos problemas. A veces, solo necesitamos un poco de información clara para entender mejor una situación. Estoy aquí para ayudarte con cualquier duda o pregunta que tengas, y siempre encontraremos una solución de manera sencilla.

GUIÓN

Bloque 1: Frases Neutrales (1-50)

1. Hoy voy a trabajar en la computadora todo el día.
2. La reunión de la tarde fue cancelada.
3. El informe de ventas será entregado mañana.
4. Hoy preparé una presentación para el equipo.
5. El auto necesita una revisión de aceite.
6. La computadora está funcionando sin problemas.
7. La conexión a internet fue interrumpida por unos minutos.
8. El proyecto debe estar terminado antes del viernes.
9. La clase de hoy fue bastante técnica.
10. Ayer hice una compra en línea.
11. El archivo se descargó correctamente.
12. Estoy organizando las carpetas en el escritorio.
13. La impresora necesita un nuevo cartucho de tinta.
14. El correo fue enviado a las nueve de la mañana.
15. Hoy tuve una reunión virtual con el equipo.

Ilustración 10: *Guión para audios del dataset*

Para asegurar que el dataset mantuviera coherencia y calidad óptima en el entrenamiento del modelo Tacotron 2, se llevó a cabo un **proceso de revisión manual exhaustivo**. Este paso fue crucial, ya que permitió identificar y corregir posibles imperfecciones en las grabaciones antes de que fueran utilizadas en el entrenamiento del sistema de síntesis de voz.

Durante esta fase, se detectaron variaciones no deseadas en el volumen, interferencias de ruido ambiental y distorsiones en la calidad del audio. De no haberse corregido, estos problemas habrían afectado negativamente la capacidad del modelo para aprender patrones fonéticos y prosódicos con precisión, comprometiendo la naturalidad e inteligibilidad de la voz sintetizada.

Para evitar estas deficiencias, se implementó un proceso de limpieza y optimización del audio, utilizando Audacity como herramienta principal para la eliminación de ruidos no deseados. Se aplicaron filtros avanzados de reducción de ruido con parámetros previamente optimizados, asegurando que las grabaciones conservaran su claridad y fidelidad fonética sin afectar la integridad del contenido original.

El impacto de este preprocesamiento pudo observarse en la Ilustración 11 y la Ilustración 12, donde se compararon las grabaciones antes y después de la aplicación de estas técnicas. La eliminación de interferencias permitió que el modelo recibiera datos más limpios y homogéneos, lo que optimizó su desempeño durante la fase de entrenamiento.

Además de mejorar la calidad del dataset, este procedimiento facilitó un entrenamiento más eficiente, reduciendo la necesidad de ajustes posteriores en la arquitectura del modelo y contribuyendo a una síntesis de voz más fluida y natural en la etapa de inferencia. La limpieza y estandarización del audio garantizaron que cada grabación utilizada en el entrenamiento

estuviera libre de distorsiones, permitiendo que el modelo aprendiera con mayor precisión las características acústicas necesarias para una conversión de texto a voz realista y expresiva.

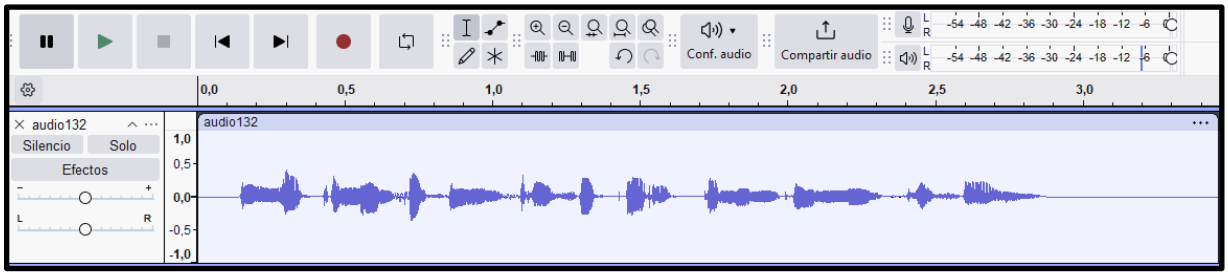


Ilustración 11: *Audio sin revisar*

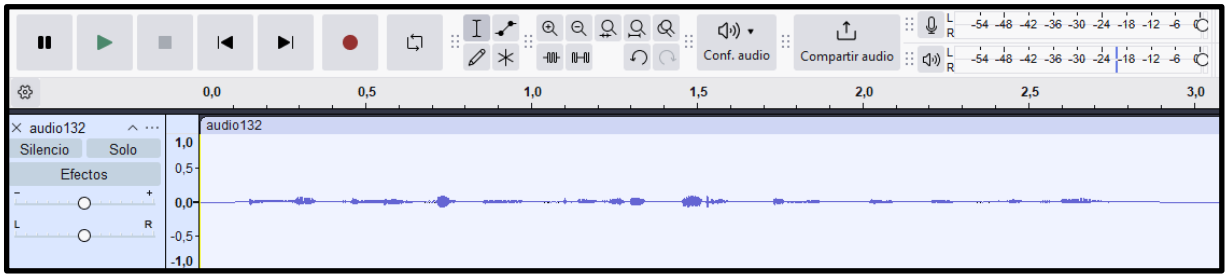


Ilustración 12: *Audio revisado*

A medida que avanzaba el desarrollo del sistema, se identificó la necesidad de ampliar significativamente el volumen de datos disponibles para el entrenamiento del modelo Tacotron 2. Contar con una mayor cantidad de muestras no solo permitió mejorar la capacidad del modelo para generalizar patrones fonéticos, sino que también ayudó a capturar una mayor variabilidad prosódica y tonal, clave para lograr una síntesis de voz más natural y expresiva.

Para abordar este desafío, se implementó una **estrategia de recolección y diversificación de datos**, enfocada en **incrementar la cantidad de audios** sin comprometer la calidad del dataset. En lugar de depender únicamente de grabaciones controladas, se adoptó un enfoque basado en la extracción de fragmentos de audiolibros provenientes de videos. Esta metodología permitió acceder a una amplia variedad de entonaciones, ritmos y estilos de habla, asegurando que el modelo pudiera aprender a generar una voz más dinámica y con mayor riqueza expresiva.

Gracias a esta estrategia, el conjunto de datos creció a 700 muestras de audio, lo que fortaleció la capacidad del modelo para capturar patrones fonéticos más complejos y mejorar su desempeño en escenarios de habla variada. Esta expansión no solo optimizó la cobertura de características lingüísticas esenciales, sino que también redujo el riesgo de sobreajuste, asegurando que el modelo no se limitara a un subconjunto restringido de datos y pudiera adaptarse mejor a distintos tipos de entrada.

Para mantener un manejo estructurado y eficiente del dataset, se implementó un sistema de nomenclatura numérica, lo que permitió mantener un esquema uniforme de almacenamiento y clasificación. Esta conversión de archivos a identificadores numéricos facilitó la organización

de los audios y su integración en el flujo de entrenamiento del modelo, evitando confusiones y mejorando la accesibilidad de los datos en cada iteración del desarrollo.

En la Ilustración 9, se presentó una muestra de los audios procesados, destacando la importancia de esta estrategia para mejorar la escalabilidad y eficiencia del sistema de conversión de texto a voz. Seguidamente, la estructuración del dataset en un formato optimizado permitió que la fase de preprocesamiento fuera más eficiente, reduciendo la complejidad en la carga de datos y asegurando que el modelo recibiera entradas acústicamente limpias, variadas y representativas del español latinoamericano. Ver ilustración 13.

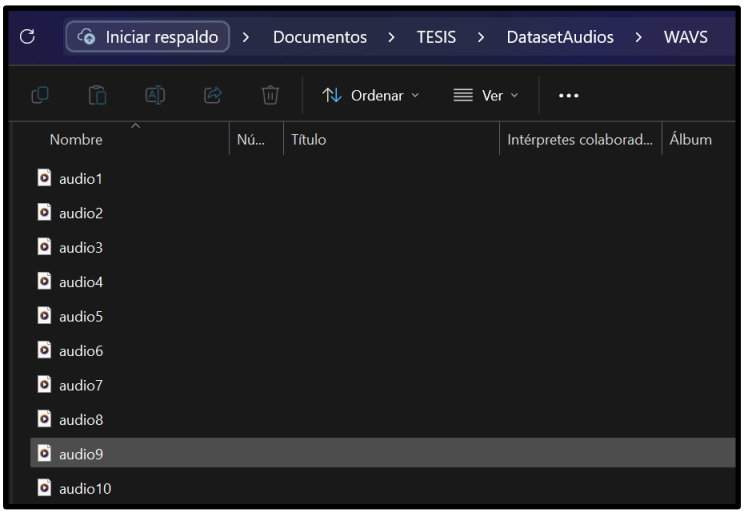


Ilustración 13: *Audios del dataset*

Dado el crecimiento del dataset y la necesidad de mantener un control estricto sobre la organización de los archivos de audio, se implementó un **sistema de renombrado secuencial** que permitió estructurar y gestionar los datos de manera eficiente. Este procedimiento fue clave para evitar duplicaciones, errores en la clasificación y posibles dificultades en la recuperación de archivos durante las etapas de entrenamiento y evaluación del modelo de síntesis de voz.

Para automatizar este proceso, se desarrolló un script en Python utilizando la biblioteca os, el cual recorrió el directorio donde se almacenaban los archivos de audio y detectó aquellos con la extensión .wav. Una vez identificados, los archivos fueron ordenados alfanuméricamente y renombrados siguiendo un índice numérico secuencial, asegurando una estructura de dataset homogénea y de fácil manejo.

Este procedimiento, demostrado en la Ilustración 13, ofreció múltiples beneficios que mejoraron la eficiencia en la manipulación del dataset,y estos beneficios son los siguientes:

- **Estandarización del dataset:** Se logró una organización uniforme de los datos, reduciendo la complejidad en su administración y facilitando su procesamiento.

- **Optimización en la carga de datos:** Al contar con nombres estructurados, la indexación y el procesamiento en lotes durante el entrenamiento del modelo se realizaron de manera más eficiente.
- **Reducción del riesgo de errores:** Se evitó el uso de nombres de archivo con caracteres especiales o espacios en blanco, que podrían haber generado problemas en la lectura automatizada de los datos.

Como se muestra en la Ilustración 13, el script permitió transformar nombres genéricos como audio1.wav, audio2.wav, etc., en un formato estandarizado (1.wav, 2.wav, etc.), simplificando la identificación y el acceso a los archivos dentro del flujo de trabajo del modelo. Esta conversión contribuyó significativamente a la automatización del preprocesamiento de datos, asegurando que cada archivo estuviera correctamente etiquetado y organizado antes de ser utilizado en el entrenamiento de Tacotron 2.

Al implementar este sistema de renombrado, se garantizó una gestión estructurada del dataset, permitiendo que el modelo de síntesis de voz pudiera acceder y procesar los datos de manera eficiente, optimizando así la calidad del entrenamiento y facilitando la integración del dataset en el pipeline de conversión de texto a voz como se observa en la Ilustración 14.

CONVERSION DE NOMBRES DE AUDIOS A NUMERICOS SECUENCIALES

```python
import os

folder_path = "audios/audios"

wav_files = [f for f in os.listdir(folder_path) if f.endswith('.wav')]

wav_files_sorted = sorted(wav_files, key=lambda x: int(''.join(filter(str.isdigit, x))))

for i, file in enumerate(wav_files_sorted, start=1):
    old_path = os.path.join(folder_path, file)
    new_file_name = f"{i}.wav"  # Solo números
    new_path = os.path.join(folder_path, new_file_name)
    os.rename(old_path, new_path)
    print(f"Renombrado: {file} -> {new_file_name}")

print("Todos los archivos han sido renombrados correctamente.")
```

```
Renombrado: audio1.wav -> 1.wav
Renombrado: audio2.wav -> 2.wav
Renombrado: audio3.wav -> 3.wav
Renombrado: audio4.wav -> 4.wav
Renombrado: audio5.wav -> 5.wav
Renombrado: audio6.wav -> 6.wav
Renombrado: audio7.wav -> 7.wav
Renombrado: audio8.wav -> 8.wav
Renombrado: audio9.wav -> 9.wav
Renombrado: audio10.wav -> 10.wav
```

Ilustración 14: *Conversión de nombres de audios*

Con el dataset correctamente estructurado y organizado, el siguiente paso en el desarrollo del sistema fue el **preprocesamiento del audio**, una fase crítica que garantizó que los datos de entrada tuvieran una calidad uniforme y estuvieran libres de distorsiones antes de ser utilizados en el entrenamiento del modelo de síntesis de voz.

Para asegurar que las grabaciones fueran lo más limpias y representativas posible, se implementaron diversas técnicas de procesamiento, optimizando la calidad acústica de los datos y facilitando su interpretación por parte del modelo Tacotron 2.

Uno de los procedimientos más importantes fue la eliminación de ruido no deseado mediante el uso de filtros digitales avanzados. Este paso resultó fundamental para reducir artefactos acústicos que podrían haber interferido con la precisión del modelo en la generación de voz sintetizada. Al minimizar el ruido de fondo y otras interferencias, se garantizó que los patrones fonéticos fueran capturados con mayor claridad, permitiendo que el modelo aprendiera de manera más eficiente.

Además, se llevó a cabo un proceso de normalización de amplitud, asegurando que todos los archivos de audio mantuvieran un nivel de volumen uniforme. Como se observa en la Ilustración 14, este procedimiento incluyó la aplicación de un valor máximo predefinido (MAX_WAV_VALUE), que permitió preservar la dinámica de la señal sin provocar saturaciones ni pérdida de información acústica relevante. Gracias a esta normalización, el volumen de las grabaciones se mantuvo estable en todo el dataset, evitando variaciones bruscas que podrían haber afectado la coherencia del entrenamiento.

Otro paso esencial en el preprocesamiento fue la transformación de las señales de audio en espectrogramas de Mel, una representación frecuencial que capturó la distribución espectral de la señal a lo largo del tiempo. Este proceso permitió que el modelo Tacotron 2 aprendiera las características acústicas del habla con mayor precisión, facilitando la generación de voz con una entonación y prosodia más naturales.

El preprocesamiento no solo mejoró la calidad de los datos utilizados en el entrenamiento, sino que también optimizó el rendimiento del modelo al proporcionarle información más estructurada y libre de ruido. Estas técnicas aseguraron que la síntesis de voz resultante tuviera una mayor claridad, estabilidad y naturalidad, acercándose cada vez más a una representación realista del habla humana. Ver ilustración 15.

```
audio_denoised = audio_denoised.cpu().numpy().reshape(-1)

normalize = (MAX_WAV_VALUE / np.max(np.abs(audio_denoised))) ** 0.9
audio_denoised = audio_denoised * normalize
wave = resampy.resample(
    audio_denoised,
    h.sampling_rate,
    h2.sampling_rate,
    filter="sinc_window",
    window=scipy.signal.windows.hann,
    num_zeros=8,
)
wave_out = wave.astype(np.int16)
```

Ilustración 15: *Preprocesamiento audios*

Seguido de esto, en la Ilustración 14, se mostró un **espectrograma de Mel** generado tras la aplicación del **preprocesamiento del audio**. En esta representación visual, las variaciones de color indicaron la intensidad de las frecuencias a lo largo del tiempo, proporcionando una referencia clave sobre la estructura fonética y acústica de la señal. Gracias a esta conversión, el modelo pudo analizar de manera más precisa los patrones de habla, lo que permitió generar una síntesis de voz con mayor claridad, fluidez y naturalidad.

Las mejoras introducidas en esta fase fueron determinantes para optimizar la calidad del dataset. La reducción significativa del ruido permitió eliminar interferencias que podrían haber afectado la inteligibilidad del habla generada, mientras que la uniformidad en el volumen garantizó que todas las grabaciones tuvieran niveles de audio consistentes, evitando distorsiones o desequilibrios que pudieran haber perjudicado el entrenamiento del modelo. Además, la transformación en espectrogramas de Mel proporcionó una representación espectral más rica y estructurada, facilitando que el modelo aprendiera con mayor precisión las características acústicas esenciales para la conversión de texto a voz.

Gracias a estas optimizaciones, el modelo Tacotron 2 fue entrenado sobre un conjunto de datos depurado y altamente representativo del español latinoamericano, lo que sentó una base sólida para la síntesis de voz. La combinación de un preprocesamiento riguroso y una correcta estructuración de los datos garantizó que el sistema pudiera generar audio de alta calidad, con entonación natural y prosodia bien definida, acercándose cada vez más a una reproducción realista del habla humana como se observa en la Ilustración 16.

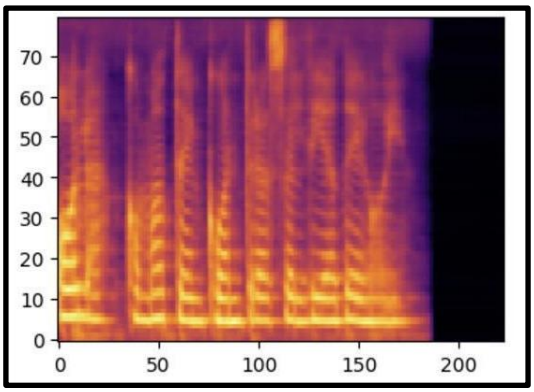


Ilustración 16: *Resultado espectrograma de Mel*

Además de haber optimizado la calidad del dataset y garantizado una representación espectral precisa a través de los espectrogramas de Mel, otro aspecto clave en el desarrollo del sistema fue la infraestructura computacional utilizada para el entrenamiento del modelo Tacotron 2. Debido a la alta demanda de procesamiento que implicaba la síntesis de voz, fue fundamental contar con aceleración por GPU, lo que permitió reducir significativamente los tiempos de entrenamiento y mejorar la eficiencia del sistema.

Para asegurar que los recursos gráficos estuvieran correctamente asignados, se llevó a cabo una **verificación y monitoreo del uso de la GPU** durante el entrenamiento del modelo. Se implementó un script en Python utilizando PyTorch, el cual permitió validar la disponibilidad de CUDA, confirmando que el entorno de ejecución era compatible con la GPU y evitando posibles errores durante la fase de entrenamiento.

Como se observa en la Ilustración 16, este proceso incluyó la consulta de varios parámetros críticos, entre ellos:

- Número de GPUs disponibles en el sistema.
- Identificación del modelo de la tarjeta gráfica, asegurando que cumpliera con los requisitos del entrenamiento.
- Memoria total asignada al dispositivo, verificando que fuera suficiente para manejar las operaciones de alto cómputo requeridas por el modelo.

Este procedimiento fue esencial para determinar si el hardware disponible contaba con soporte adecuado para la ejecución del modelo en GPU. La compatibilidad con CUDA resultó crucial, ya que permitió realizar cálculos matriciales intensivos de manera eficiente, acelerando la generación de espectrogramas de Mel y mejorando el rendimiento en la inferencia del modelo de síntesis de voz.

Gracias a esta validación, se garantizó que el sistema operara en un entorno óptimo y optimizado, asegurando que los procesos de conversión de texto a voz se ejecutaran con la menor latencia posible y con un uso eficiente de los recursos computacionales. Ver ilustración 17.

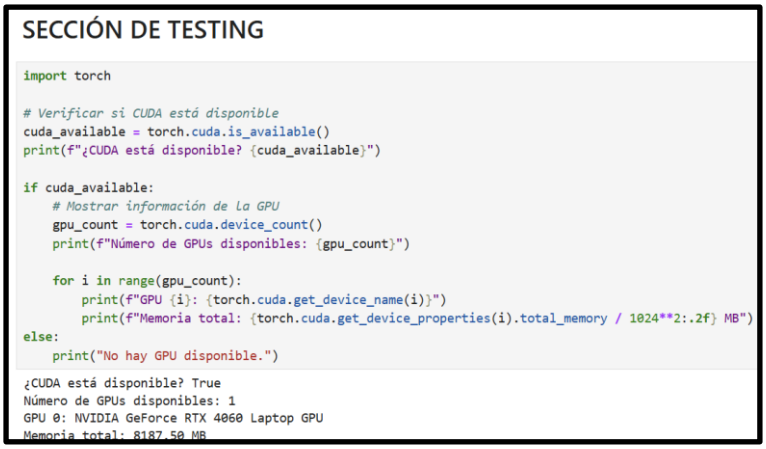


Ilustración 17: *Visualización de GPU*

Con la disponibilidad de la GPU confirmada y la compatibilidad con CUDA validada, el siguiente paso fue la **implementación de modelos preliminares para evaluar el rendimiento del sistema**. Esta fase permitió generar los primeros audios de síntesis de voz en español latinoamericano, asegurando que la integración de los modelos en el entorno de desarrollo fuera funcional y que los resultados iniciales cumplieran con los estándares esperados en términos de naturalidad, fluidez e inteligibilidad.

Para esta evaluación, se probaron distintos modelos de conversión de texto a voz, incluyendo Tacotron2-DDC, Mozilla TTS y Bark. Cada uno de estos modelos fue ejecutado con diferentes configuraciones para analizar su desempeño y comparar la calidad del audio sintetizado.

En la Ilustración 17, se presentó un fragmento del audio generado, donde se pudo observar el proceso de síntesis exitoso. En esta representación, se mostró la ruta de salida del archivo de

audio y la velocidad de procesamiento, dos factores clave para medir la eficiencia del sistema en la conversión de texto a voz.

Los resultados obtenidos en esta prueba preliminar validaron la correcta integración de los modelos TTS en el entorno de desarrollo, permitiendo comprobar la calidad inicial de las voces sintetizadas. Estos audios sirvieron como punto de referencia para realizar ajustes posteriores, optimizando parámetros como entonación, claridad y estabilidad del habla generada, con el objetivo de lograr una síntesis de voz más natural y expresiva. Ver Ilustración 18.

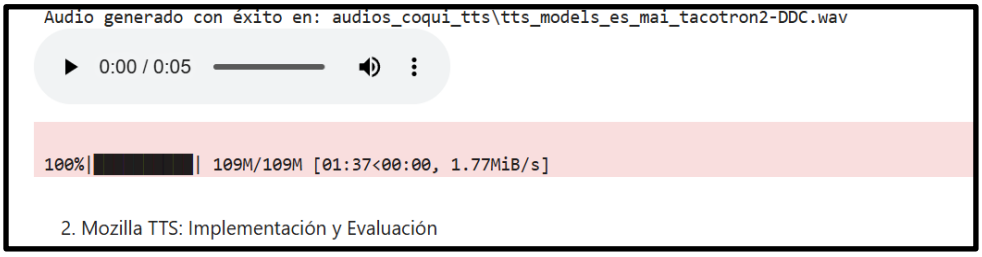


Ilustración 18: *Audio generado*

El uso de una GPU no sólo permitió la implementación exitosa de los modelos de síntesis de voz, sino que también aceleró significativamente el proceso de entrenamiento. La conversión de texto a voz fue un procedimiento computacionalmente intenso, ya que implicó la manipulación de grandes volúmenes de datos y la ejecución de múltiples operaciones matriciales en paralelo. Sin esta optimización, el entrenamiento habría requerido tiempos computacionales considerablemente más largos, afectando la viabilidad y escalabilidad del desarrollo del sistema.

Gracias a la integración eficiente del hardware, el modelo pudo procesar y analizar los espectrogramas de Mel en menos tiempo, permitiendo ajustes más rápidos en los parámetros de entrenamiento y logrando una convergencia más estable. Además, la aceleración por GPU facilitó la experimentación con diferentes configuraciones del modelo, optimizando tanto la calidad del audio sintetizado como el rendimiento en la inferencia.

Estas configuraciones aseguraron que la arquitectura del modelo operará de manera eficiente, maximizando el uso del hardware disponible para lograr una síntesis de voz en tiempo real, sin comprometer la naturalidad ni la fluidez del habla generada. La combinación de una infraestructura bien optimizada y modelos entrenados con datos de alta calidad permitió consolidar un sistema de conversión de texto a voz capaz de generar audio con una entonación más precisa y expresiva.

Con la infraestructura computacional optimizada y la validación del uso de GPU y CUDA, el siguiente paso en el desarrollo del sistema fue la **implementación y entrenamiento inicial del modelo Tacotron 2.** En esta tercera iteración, el enfoque se centró en ajustar la configuración del modelo para garantizar una síntesis de voz clara, fluida y natural, optimizando los parámetros clave que influyeron en su estabilidad y desempeño durante el entrenamiento.

Para lograrlo, se definieron hiperparámetros esenciales, los cuales regularon el comportamiento del modelo en cada etapa de su aprendizaje. La Ilustración 18 presenta la configuración utilizada en esta fase, destacando los parámetros más relevantes y su impacto en el proceso:

- **Tasa de aprendizaje (hparams.A_ = 3e-4):** Se estableció en 0.0003 (3e-4) para garantizar un ajuste controlado de los pesos del modelo. Una tasa muy alta habría provocado inestabilidad en la convergencia, mientras que una demasiado baja habría ralentizado excesivamente el aprendizaje.
- **Carga de espectrogramas desde disco (hparams.load_mel_from_disk = True):** Se habilitó esta opción para evitar el recálculo innecesario de los espectrogramas de Mel en cada iteración, optimizando el tiempo de procesamiento y reduciendo la carga computacional.
- **Limpieza de texto (hparams.text_cleaners = ["basic_cleaners"]):** Se utilizó el método "basic_cleaners" para preprocesar el texto de entrada, eliminando caracteres innecesarios y mejorando la correspondencia entre el texto y el audio generado.
- **Capas ignoradas en el entrenamiento (hparams.ignore_layers = ["embedding.weight"]):** Se excluyeron ciertas capas del modelo preentrenado para preservar representaciones embebidas relevantes, permitiendo una mejor adaptación sin comprometer la información aprendida previamente.
- **Número de épocas (hparams.epochs = 250):** Se definieron 250 épocas, asegurando que el modelo alcanzara un nivel de aprendizaje adecuado sin caer en sobreajuste.
- **Intervalo de guardado (saving_interval = 1):** Se configuró el guardado de los pesos del modelo en cada época, lo que permitió recuperar versiones intermedias en caso de ser necesario y realizar ajustes en tiempo real.
- **Optimización con CUDA (torch.backends.cudnn.enabled = hparams.cudnn_enabled):** Se habilitó el uso de cuDNN (CUDA Deep Neural Network) para mejorar la eficiencia del entrenamiento en GPU, reduciendo considerablemente los tiempos de cómputo.
- **Optimización de rendimiento (torch.backends.cudnn.benchmark = hparams.cudnn_benchmark):** Se activó esta configuración para permitir que cuDNN seleccionara automáticamente la mejor estrategia de ejecución según el hardware disponible, maximizando la velocidad del entrenamiento.

El ajuste de estos hiperparámetros fue clave para asegurar que el modelo aprendiera de manera eficiente, optimizando la conversión de texto a voz sin perder estabilidad. Una correcta gestión de la tasa de aprendizaje y el número de épocas permitió alcanzar un equilibrio entre la velocidad de convergencia y la calidad del audio sintetizado, evitando entrenamientos prolongados que no aportaran mejoras significativas.

Además, la implementación de técnicas avanzadas de optimización con cuDNN y CUDA permitió reducir drásticamente los tiempos de entrenamiento, haciendo viable la ejecución de Tacotron 2 en hardware especializado. La combinación de una arquitectura bien ajustada y una

infraestructura optimizada sentó las bases para que el modelo generara audios de alta calidad, con una entonación natural y una prosodia bien definida.

En la Ilustración 17, se ilustra la configuración utilizada en esta fase, proporcionando una referencia visual sobre los parámetros que regularon el comportamiento de la red neuronal y su impacto en la síntesis de voz. Ver Ilustración 19.

```
  hparams.A_ = 3e-4

hparams.load_mel_from_disk = True
hparams.text_cleaners=["basic_cleaners"]
hparams.ignore_layers=["embedding.weight"]

hparams.epochs =  250


saving_interval =  1

torch.backends.cudnn.enabled = hparams.cudnn_enabled
torch.backends.cudnn.benchmark = hparams.cudnn_benchmark
```

Ilustración 19: *Configuración de parámetros*

Con la configuración del modelo establecida y los parámetros ajustados, se procedió a la fase de **entrenamiento preliminar,** en la cual se llevaron a cabo las primeras iteraciones para evaluar el desempeño de Tacotron 2 en la conversión de texto a voz. Durante esta etapa, se procesaron los datos del dataset optimizado, permitiendo la generación de los primeros espectrogramas de Mel, una representación clave que reflejaba la transformación de las características fonéticas del texto en señales acústicas que posteriormente fueron convertidas en audio.

En la Ilustración 19, se muestra uno de los primeros resultados obtenidos tras el entrenamiento inicial. Esta imagen permitió analizar tres elementos fundamentales que dieron indicios del estado del modelo en sus primeras etapas:

- **Analisis del espectrograma de Mel (izquierda)**
    - Se observó la representación espectral de la voz sintetizada, donde la escala vertical representaba las frecuencias en Hz y el eje horizontal indicaba la progresión temporal del audio.
    - La intensidad de los colores reflejaba la energía en cada banda de frecuencia: los tonos más brillantes indicaban mayor amplitud, mientras que las áreas más oscuras representaban regiones de menor intensidad.
    - Se logró una generación inicial de estructuras armónicas, aunque aún se evidenciaban ciertos defectos en la transición entre fonemas, lo que sugería la necesidad de ajustes en la configuración del modelo.
- **Análisis de alineación (derecha)**

    - Se presentó la alineación entre el texto y las características acústicas generadas por el modelo.
    - La diagonalidad en la gráfica indicó que la red neuronal logró mapear la secuencia de texto a audio de manera consistente, aunque con algunas desviaciones menores que requerían refinamiento en los hiperparámetros para mejorar la precisión.
- **Audio sintetizado (parte inferior)**
    - Se generó una muestra de voz basada en el espectrograma de Mel, la cual fue evaluada por los investigadores.
    - A partir de esta prueba, se identificó que, si bien la entonación y la claridad de las palabras eran aceptables, aún era necesario refinar la prosodia y mejorar la fluidez en las pausas y transiciones fonéticas, ya que algunas frases sonaban mecánicas o con cortes abruptos.

Estos primeros resultados permitieron validar el funcionamiento del modelo y detectar áreas que requerían ajustes adicionales antes de proceder con un entrenamiento más extenso. Con base en estas observaciones, se planificaron nuevas optimizaciones en los hiperparámetros y en la estructura del modelo para mejorar la naturalidad y estabilidad del habla sintetizada en las siguientes iteraciones. Ver Ilustración 20.

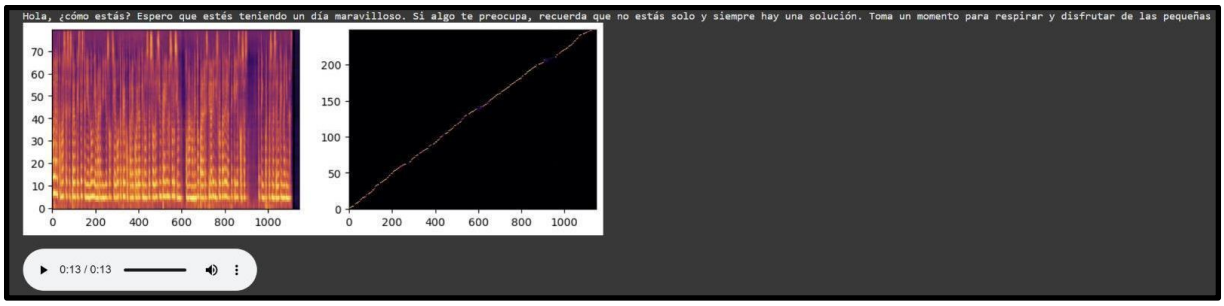


Ilustración 20: *Primeros resultados de espectrograma de Mel*

El análisis de estos espectrogramas preliminares permitió identificar áreas clave que requerían mejoras significativas en la síntesis de voz. Aunque el modelo logró generar estructuras armónicas y mapeó el texto a audio de manera consistente, aún presentaba deficiencias que debían ser corregidas antes de avanzar a un entrenamiento más profundo.

Las principales observaciones incluyeron:

- **Deficiencias en la prosodia:**
    - La voz sintetizada carecía de variabilidad emocional y no presentaba una entonación natural, lo que resultaba en una pronunciación monótona y poco expresiva. Esta limitación indicó la necesidad de ampliar el volumen de datos en el dataset, incorporando grabaciones con distintas expresiones y tonalidades para mejorar la capacidad del modelo de reproducir un habla más natural.
- **Consumo elevado de memoria:**

    - Se observó un uso intensivo de recursos computacionales, lo que afectaba el rendimiento del sistema. Para optimizar el proceso de entrenamiento, se planificó un ajuste en la configuración de GPU, asegurando una mejor gestión de la memoria y reduciendo los tiempos de cómputo sin comprometer la calidad del modelo.
- **Corrección en la síntesis de fonemas:**
    - En algunos casos, los fonemas presentaban discontinuidades o alteraciones en su duración, generando inconsistencias en la fluidez del habla sintetizada. Esta anomalía sugirió la necesidad de ajustes en la arquitectura del modelo, así como la posibilidad de modificar la cantidad de iteraciones de entrenamiento para mejorar la estabilidad en la generación de sonidos.

Como resultado, la evaluación de estos espectrogramas proporcionó información crucial para la optimización del modelo, permitiendo ajustar los hiperparámetros y definir estrategias específicas para mejorar su desempeño en iteraciones posteriores.

En la Ilustración 19, se pudo corroborar visualmente el progreso inicial del modelo, facilitando la detección de patrones de error y áreas de mejora, lo que permitió diseñar una estrategia más efectiva para refinar la síntesis de voz en las siguientes fases del entrenamiento.

Con base en los resultados obtenidos en iteraciones anteriores, la siguiente fase se centró en el **refinamiento y evaluación del modelo**, con el objetivo de mejorar la calidad y naturalidad de la síntesis de voz. Durante este sprint, se realizaron ajustes en los hiperparámetros clave, optimizando aspectos como la prosodia, la entonación y la fluidez en las transiciones entre fonemas.

Para lograr una mayor expresividad en el audio generado, se implementaron optimizaciones específicas en la configuración del modelo, las cuales son las siguientes:

- **Ajuste de la tasa de aprendizaje:** Se modificaron los valores de aprendizaje para mejorar la estabilidad del entrenamiento, asegurando que el modelo pudiera converger a una síntesis de voz más precisa y natural.
- **Optimización de la arquitectura:** Se ajustaron ciertas capas de la red neuronal para permitir una mejor representación de los patrones acústicos del español latinoamericano.
- **Mejora en la prosodia y la entonación emocional:** Se realizaron modificaciones en los pesos de la red para que el modelo generara una voz más expresiva, evitando un tono robótico y monótono.

Estos cambios permitieron que la síntesis de voz tuviera mayor fluidez y coherencia en la transición entre fonemas, mejorando la percepción de naturalidad en las pruebas de validación.

Para evaluar el desempeño del modelo refinado, se **generaron y analizaron espectrogramas de Mel** obtenidos a partir del dataset procesado. Estos espectrogramas sirvieron para comparar

la fidelidad del audio sintetizado con respecto al audio original y detectar posibles áreas de mejora.

El análisis se centró en los siguientes aspectos:

- **Estabilidad de la señal:** Se verificó que la síntesis de voz no presentara cortes abruptos ni distorsiones inesperadas.
- Coherencia en la variación tonal: Se examinó que la modulación del tono fuera consistente con la entonación natural del habla humana.
- **Precisión en la representación de transiciones fonéticas:** Se observó cómo el modelo generaba transiciones suaves entre fonemas, sin introducir sonidos artificiales o erráticos.

En la Ilustración 20, se presenta un espectrograma de Mel correspondiente a uno de los audios generados tras los ajustes finales.

- **A la izquierda de la imagen:** Se mostró la distribución de frecuencias en función del tiempo, lo que permitió visualizar la progresión temporal del habla sintetizada.
- **A la derecha:** Se observó un gráfico de alineación, que evidenció la correspondencia entre los fonemas de entrada y la representación acústica generada por el modelo.

El análisis confirmó que las optimizaciones mejoraron la fluidez y expresividad del audio, aunque aún hay áreas para mejorar la modulación de la prosodia. Esto sugiere la necesidad de seguir ajustando el modelo en futuras iteraciones para lograr una síntesis de voz más natural y dinámica, como se muestra en la Ilustración 21

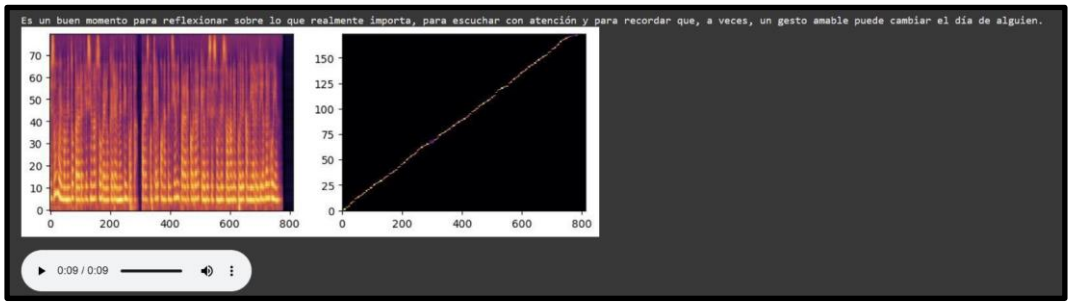


Ilustración 21: *Espectrogramas del dataset procesado*

5.4 Ejecución del plan de experimentación (OE3)

Para evaluar la calidad del modelo y su impacto en el público objetivo, se llevó a cabo una encuesta dirigida a maestras de la **Universidad Politécnica Salesiana**, quienes, además de su labor académica, trabajaban con niños pequeños. Su perspectiva fue clave porque no solo podían darnos su opinión como profesionales de la enseñanza, sino que también entendían cómo

los niños percibían una voz sintética... qué les llamaba la atención y qué podía hacer que perdieran el interés.

Durante la evaluación, compartieron observaciones muy valiosas. En general, coincidieron en que la voz generada era clara y comprensible, pero hubo algo en lo que insistieron: para los niños, podía resultar monótona. Nos recomendaron hacerla más dinámica, con una entonación más expresiva, que tuviera más vida… evitando que sonara demasiado plana o aburrida.

Para estructurar mejor los resultados, aplicamos una encuesta donde calificaron distintos aspectos del modelo con una escala del 1 al 5; donde 1 significaba "muy malo" y 5 "excelente". Gracias a esto, obtuvimos tanto puntos fuertes como áreas de mejora, lo que nos dio una idea más clara de cómo optimizar la expresividad de la voz para hacerla más atractiva en su aplicación con *Volk*, el robot.

Esta encuesta fue respondida por profesionales con experiencia en educación infantil y universitaria, incluyendo a **Ordoñez Vásquez María Elisa (Anexo 2.1), Reyes Quezada Karina Priscila (Anexo 2.2), Almeida Soliz Nidia Elizabeth (Anexo 2.3) y Zamora Torres Johana Elizabeth (Anexo 2.4)**. Su conocimiento en pedagogía y su experiencia directa con niños hicieron que sus aportes fueran fundamentales para entender cómo mejorar la voz generada y hacerla más adecuada para la interacción con un asistente robótico.

El cuestionario se estructuró en tres bloques. Primero, se recopilaron datos demográficos para conocer el perfil de las participantes. Luego, se incluyeron preguntas en escala de Likert para evaluar la voz generada en distintos aspectos técnicos y perceptivos. Al final de estas encuestas se añadió una pregunta abierta para que las maestras pudieran compartir sus opiniones y sugerencias de mejora.

VI. Resultados.

A continuación, se presentan los hallazgos obtenidos tras la implementación del sistema de conversión de texto a voz. La evaluación se centró en la precisión, naturalidad e inteligibilidad del habla sintetizada, comparando el desempeño del modelo inicial, el modelo final y los datos de referencia. Para ello, se analizaron espectrogramas, formas de onda y métricas acústicas, junto con la percepción de los usuarios en encuestas aplicadas a expertos en el área. Estos análisis permiten verificar la eficacia del modelo final y su cumplimiento con los objetivos planteados.

6.1 Evaluación de modelos de aprendizaje profundo para TTS (OE1)

Para la selección del modelo de conversión de texto a voz, se evaluaron diversas arquitecturas con el fin de determinar cuál ofrecía la mejor síntesis de voz en términos de fluidez, estabilidad y calidad. Como se observa en la **Ilustración 1**, el análisis de los espectrogramas permitió

comparar las diferencias en la segmentación de fonemas, transiciones de sonido y coherencia en la entonación entre los modelos evaluados.

Tacotron 2 mostró una distribución espectral más uniforme, transiciones más suaves y una segmentación clara de fonemas en comparación con MozillaTTS y CoquiTTS. En contraste, MozillaTTS presentaba variaciones irregulares y pausas abruptas, mientras que CoquiTTS generaba cortes más marcados en la energía de las frecuencias, afectando la fluidez del habla.

Estos hallazgos confirman que Tacotron 2 logró una síntesis de voz más estable y natural, evitando los problemas de discontinuidad y variabilidad encontrados en los otros modelos. Su capacidad para generar transiciones bien definidas y mantener una modulación consistente fue clave para su selección como el modelo más adecuado para este proyecto, cumpliendo así con el OE1.

6.2 Implementación del módulo de conversión de texto a voz (OE2)

Para evaluar la efectividad del sistema de conversión de texto a voz, se analizaron diversas métricas acústicas que permitieron medir la naturalidad, fluidez y estabilidad del habla generada. La evaluación se basó en tres análisis principales: la forma de onda, el espectrograma y la energía en el tiempo. A través de estas comparaciones, se identificaron mejoras clave logradas en el modelo final en relación con la versión inicial.

El análisis de la forma de onda reflejó cómo la segmentación del habla impacta la fluidez del audio sintetizado. Como se observa en la **Ilustración 22**, el audio original presentó pausas bien definidas y una prosodia natural, mientras que el modelo inicial mostró señales más compactas, con transiciones abruptas y pausas irregulares, lo que generaba un habla acelerada y poco natural. En contraste, el modelo final corrigió estos problemas, logrando una estructura más equilibrada y pausas mejor distribuidas. A pesar de ello, aún se identificaron ligeras diferencias en la duración de ciertas pausas, lo que sugiere oportunidades de mejora en la segmentación temporal de las frases.

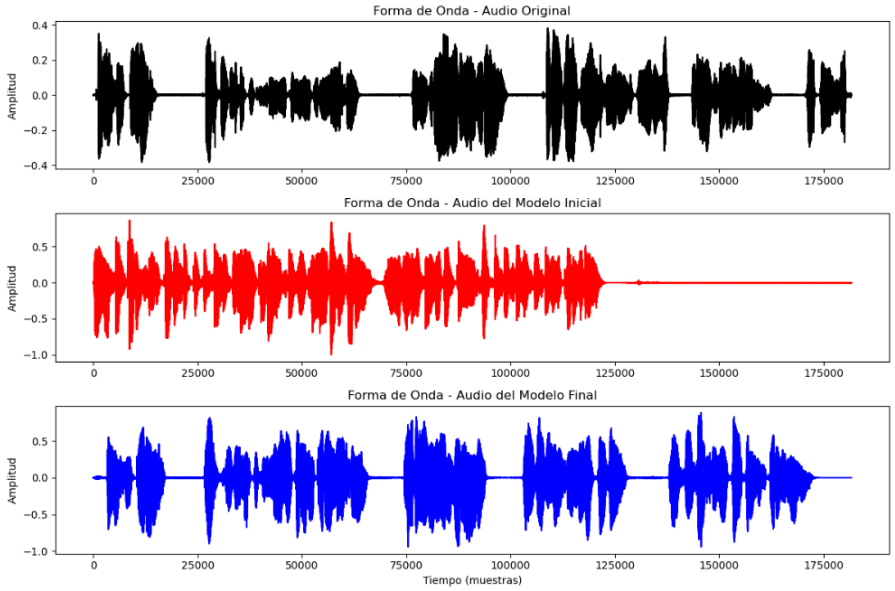


Ilustración 22: *Forma de Onda de los audios (waveform)*

Por otro lado, el análisis espectral permitió evaluar la distribución de frecuencias en el tiempo, identificando diferencias clave en la síntesis de voz. Como se muestra en la **Ilustración 23**, el audio original mantuvo bandas espectrales bien definidas y transiciones suaves entre fonemas, reflejando una prosodia natural. En cambio, el modelo inicial presentó pérdida de energía en frecuencias altas y cortes abruptos en la señal, lo que generaba una voz robótica con menor claridad. En el modelo final, estas deficiencias fueron corregidas en gran medida, logrando una continuidad espectral más uniforme y reduciendo las interrupciones abruptas. Si bien todavía se observan pequeñas pérdidas en las frecuencias altas, la mejora general permitió acercarse significativamente al comportamiento del audio original.

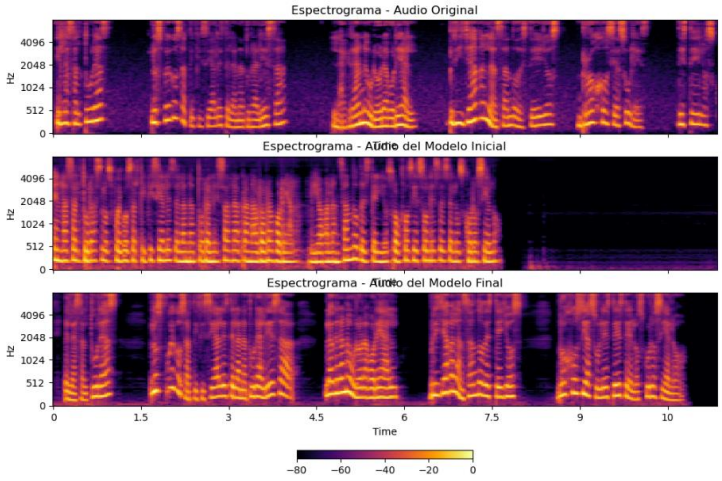


Ilustración 23: *Comparación de los espectrogramas*

Además, el análisis de la energía en el tiempo fue fundamental para medir la estabilidad de la síntesis de voz. Como se detalla en la **Ilustración 24**, el audio original mostró una modulación

equilibrada y bien distribuida, reflejando un flujo natural del habla. Sin embargo, el modelo inicial evidenció fluctuaciones abruptas y caídas repentinas en la intensidad, lo que resultó en una síntesis inestable con pronunciaciones desbalanceadas. El modelo final corrigió este comportamiento, logrando una curva de energía mucho más estable y uniforme. A pesar de estas mejoras, se identificaron algunos picos inesperados de intensidad, lo que sugiere que aún es posible optimizar la modulación de la energía para evitar pronunciaciones forzadas o exageradas.

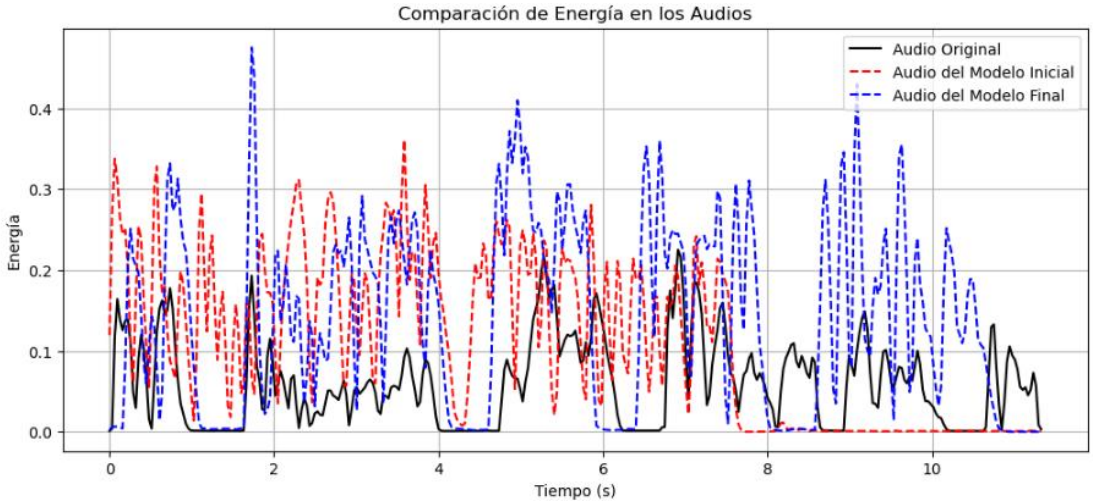


Ilustración 24: *Comparación de Energía*

Estos resultados confirman que el sistema logró cumplir con el **OE2**, ya que las optimizaciones implementadas permitieron mejorar la segmentación del habla, estabilizar la energía en el tiempo y reducir la robótica en la entonación. Si bien aún existen oportunidades para refinar ciertos aspectos, la síntesis de voz generada por el modelo final se acerca mucho más a la naturalidad del habla humana, sentando una base sólida para futuras optimizaciones.

6.3 Análisis estadístico aplicado (OE3)

Con el objetivo de evaluar la percepción de expertos respecto al diseño y desarrollo de un módulo de conversión de texto a voz (TTS) para el asistente robótico Volk, se realizó una evaluación con 4 expertas especializadas en psicología, educación infantil, educación inicial y preparatoria. Las expertas cuentan con una experiencia profesional promedio de 20 años (DS = 8.60) y una edad promedio de 41 años (DS = 6.00), donde DS representa la desviación estándar, medida que refleja la variabilidad de los datos respecto a su media. El estudio busca determinar si las características de la voz sintética, como su claridad, naturalidad y prosodia, son adecuadas para mejorar la interacción y el aprendizaje en contextos educativos inclusivos. **Ver Tabla 1.**

Experta	Apellidos y Nombres	Género	Edad	Profesión	Años de experiencia Profesional	Área de Docencia

1	Ordoñez Vásquez María Elisa	Femenino	44	Psicología Educativa/Docente	20	Inclusión Educativa
2	Reyes Quezada Karina Priscila	Femenino	32	Docente en Educación Inicial y Preparatoria	9	Educación Inicial
3	Almeida Soliz Nidia Elizabeth	Femenino	44	Licenciatura en Ciencias de la Educación, mención Educación infantil	21	Educación Inicial y Preparatoria
4	Zamora Torres Johana Elizabeth	Femenino	44	Maestría en intervención y Educación Inicial	30	Educación Inicial y Universitaria

Tabla 1. *Características del perfil de las expertas.*

Los datos recopilados mediante una encuesta aplicada a las cuatro expertas, compuesta por 26 preguntas (9 de tipo demográfico, 16 en escala de Likert y 1 pregunta abierta), se analizaron utilizando el lenguaje de programación R (versión 4.4.2) y el coeficiente de concordancia de Kendall (W). La escala de Likert, con respuestas fijas, permitió evaluar las actitudes y opiniones de las expertas, mientras que la pregunta abierta complementa la recopilación de información cualitativa.

- ✓ ***Datos demográficos:***

  Las preguntas de carácter demográfico nos permitieron recolectar la siguiente información: nombres y apellidos, género, edad, lugar de trabajo, profesión, años de experiencia laboral, si en la actualidad ejerce la docencia, su experiencia como docente con niños de 6 a 8 años y, actualmente el área en la que cumple sus labores

- ✓ ***Para la escala de Likert***, se analizaron como primera parte las cuatro sesiones del cuestionario que corresponde a las 16 preguntas, y como segunda parte del análisis la pregunta 17 de ámbito opcional y redacción personal se presenta en la tabla 4. A continuación, se plantean los resultados de las cuatro sesiones en la tabla 3.

- ✓ La pregunta 17, ***como tercera parte de la encuesta*** que analiza ¿Qué aspectos considera que podrían mejorarse en la voz generada para adaptarse mejor a la edad objetivo? A continuación, la tabla 2.

Experta	Consideraciones de la pregunta 17
Experta 1. Ordoñez Vásquez María Elisa	✓ La tonalidad de la voz en cada palabra.

Experta 2. Reyes Quezada Karina Priscila	✓ Más pausa y mayor entonación.
Experta 3. Almeida Soliz Nidia Elizabeth	✓ Tener un ritmo que varíe en función de cada temática. La entonación de palabras.
Experta 4. Zamora Torres Johana Elizabeth	✓ La voz debe ser más natural, que llame la atención de los niños, aun son pequeños y es necesario motivarlos.

Tabla 2. *Consideraciones por parte de expertas*

Las preguntas de carácter demográfico permitieron recolectar información sobre las expertas, incluyendo sus nombres y apellidos, género, edad, lugar de trabajo, profesión, años de experiencia laboral, si son docentes en la actualidad y el área en la que desempeñan sus labores. Las preguntas en escala de Likert se enfocaron en evaluar aspectos clave de la voz generada por el módulo de texto a voz (TTS), como la naturalidad de la pronunciación de las palabras (P1), la entonación de la voz para transmitir emociones positivas (P2), la claridad de la voz para facilitar la comprensión (P3), el ritmo y la entonación para lograr fluidez y expresividad (P4), la efectividad de la voz para captar la atención de los niños durante las actividades (P5), la utilidad de la voz para motivar a los niños en actividades interactivas (P6), la percepción de amigabilidad de la voz para niños de 6 a 8 años (P7), la utilidad de la voz en actividades recreativas como cuentos interactivos (P8), la calidad general de la voz en comparación con sistemas TTS estándar (P9), la adecuación de la voz para niños en comparación con otros sistemas TTS (P10), la adecuación de la entonación para captar el interés de los niños (P11), la relevancia de ajustar la prosodia para mejorar la interacción con niños de 6 a 8 años (P12), la utilidad de la voz en actividades de apoyo educativo infantil (P13), la posibilidad de usar el sistema en el hogar para reforzar el aprendizaje (P14), la importancia de la voz en aplicaciones tecnológicas educativas (P15), y la adecuación de la voz para promover la colaboración entre niños (P16). La pregunta abierta complementó la recolección de información cualitativa, permitiendo una visión más profunda de las percepciones de las expertas.

Para determinar el nivel de acuerdo entre las valoraciones de las expertas, se realizó un análisis estadístico utilizando el coeficiente de concordancia de Kendall (W). Considerando que es una herramienta muy útil para la validación de instrumentos y escalas de medición. Además, el coeficiente de Kendall plantea el grado de asociación de las evaluaciones ordinarias desarrolladas por evaluadores múltiples en situaciones que se evalúa las mismas muestras. De esta manera, los valores del coeficiente de Kendall tienen un rango de 0 a +1 (Maskavizan, A. J., Poco, A. N., & Calzolari, 2023). Con estos antecedentes, este análisis permitió medir la consistencia en las respuestas de las expertas respecto a los aspectos evaluados. Se plantearon las siguientes hipótesis:

H_0: No existe consistencia en las valoraciones de las expertas para los criterios evaluados.
H_a: Sí existe consistencia en las valoraciones de las expertas para los criterios evaluados.

De acuerdo con la interpretación propuesta por la Ilustración 25, el valor calculado para el coeficiente de concordancia de Kendall (W = 0.703) sugiere un nivel de acuerdo "fuerte" entre

las expertas. El valor de p asociado (p = 0.001) indica que se puede rechazar la hipótesis nula, lo que significa que existe evidencia suficiente para afirmar que las valoraciones de las expertas presentan una consistencia significativa. Este resultado respalda la efectividad del módulo de texto a voz en términos de claridad, naturalidad y adecuación para contextos educativos infantiles.

Interpretación del coeficiente de Kendall (W):

- $0.00 \leq W < 0.20$: Acuerdo leve.
- $0.20 \leq W < 0.40$: Acuerdo moderado.
- $0.40 \leq W < 0.60$: Acuerdo sustancial.
- $0.60 \leq W < 0.80$: Acuerdo fuerte.
- $W \geq 0.80$: Acuerdo casi perfecto.

El valor de **W = 0.703** indica un acuerdo fuerte entre las expertas, lo que refleja una percepción positiva y consistente respecto a la calidad de la voz generada. Las expertas destacaron la naturalidad, claridad y adecuación de la voz para aplicaciones educativas, especialmente en la interacción con niños de 6 a 8 años. Además, se resaltó la importancia de ajustar la prosodia y el ritmo para mejorar la expresividad y el interés de los niños en actividades interactivas. Estos resultados sugieren que el módulo de texto a voz desarrollado cumple con los objetivos de naturalidad, empatía y claridad requeridos para su implementación en entornos educativos inclusivos.

En la Ilustración 25 se pueden observar cuatro diagramas aluviales que reflejan la percepción de cuatro expertas, considerando su experiencia laboral. En el cuadrante A, que evalúa la claridad de la voz generada, se aprecia que, si bien hay diferencias en la percepción, la mayoría coincide en que tiene un nivel moderado. Esto sugiere que la prosodia del sistema es comprensible, aunque puede mejorar en nitidez para una mejor experiencia auditiva. De manera similar, en el cuadrante C, que analiza la adecuación para niños, se observa una tendencia positiva, donde a mayor experiencia laboral, mayor es la percepción de que el sistema es apropiado para este público, destacando su potencial en entornos educativos.

Por otro lado, los cuadrantes B y D, relacionados con la utilidad en actividades recreativas y la importancia en aplicaciones educativas, muestran una evolución favorable en la percepción del sistema. A medida que aumenta la experiencia de las expertas, también crece la valoración positiva, alcanzando niveles altos y muy altos. Esto indica que el sistema tiene un gran potencial tanto en el ámbito lúdico como en el educativo, con oportunidades de mejora en aspectos específicos como la prosodia y la claridad de la voz. Ver Ilustración 25.

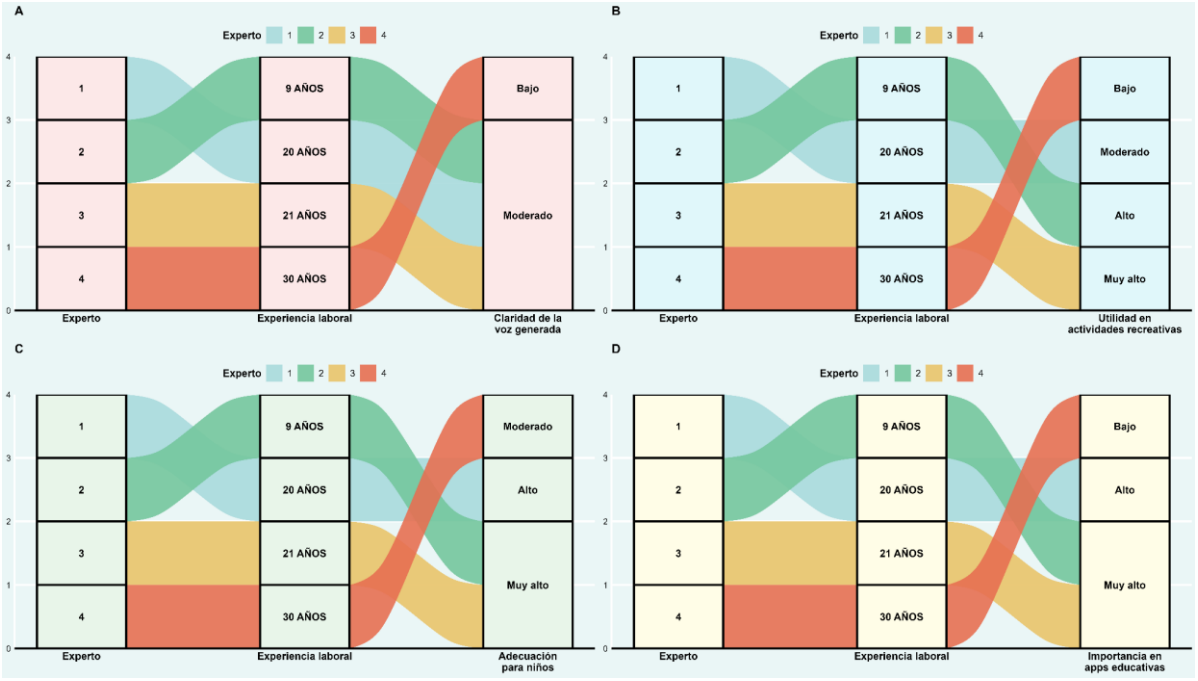


Ilustración 25: *Resultados del análisis de consenso*

En la Ilustración 26 se muestra un resumen de la coincidencia entre los criterios de las cuatro expertas en las diferentes preguntas: a) consenso total entre las cuatro expertas (P1, P2, P3, P5, P6, P8, P10, P11, P12 y P16), b) consenso parcial de tres expertas en tres preguntas (P4, P7 y P14), c) consenso parcial de dos expertas en dos preguntas (P9 y P15), y d) ausencia de consenso en ninguna pregunta.

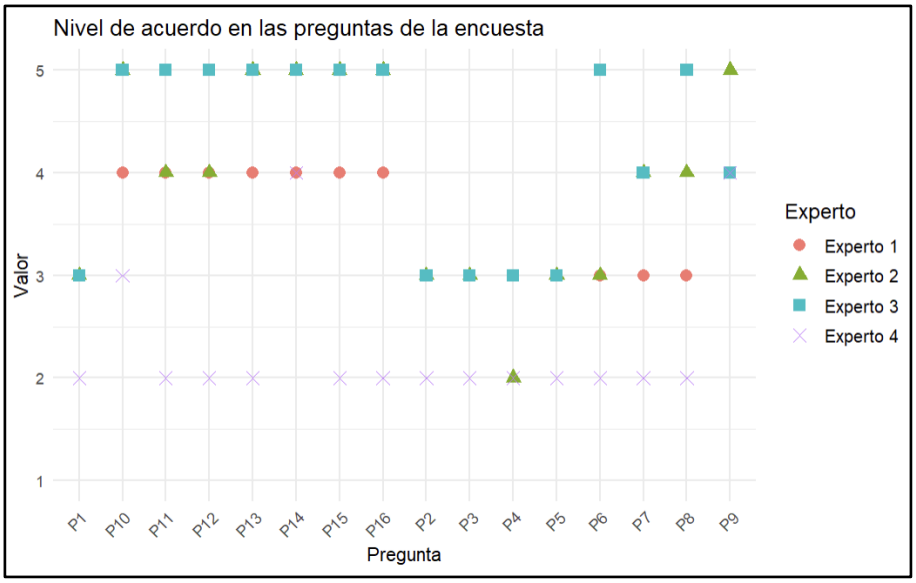


Ilustración 26: *Nivel de coincidencia en las respuestas de las expertas.*

En la primera parte de la encuesta, se destacó la importancia del nivel de prosodia, mientras que en la segunda parte se obtuvieron resultados relevantes en relación con el nivel de utilidad del sistema. Además, la sesión aleatoria proporcionó sugerencias clave, como ajustes en la

entonación, que serán consideradas en las recomendaciones para mejorar el diseño del módulo en futuras investigaciones. De esta manera, Se identificó un nivel de aceptación superior al 70%, lo que sugiere que el modelo de conversión de texto a voz es eficiente y aplicable para niños de 6 a 8 años, con posibilidades de implementación en asistentes robóticos educativos.

VII. Cronograma

Para desarrollar el cronograma de trabajo se consideró los objetivos planteados con la finalidad de obtener los resultados esperados; a continuación, las actividades por cada objetivo:

- OE1: Estudiar y evaluar las principales redes neuronales de aprendizaje profundo y herramientas para la generación de texto a voz, considerando aspectos de entonación y prosodia, mediante la realización de pruebas comparativas.

No.	**Actividad**
1	Investigación sobre redes neuronales de aprendizaje profundo (RNN, LSTM, Transformer)
2	Análisis de modelos preentrenados para TTS (Tacotron, Coqui TTS, etc.)
3	Pruebas iniciales con modelos pre entrenados y evaluación de compatibilidad con español de Latinoamérica.
4	Evaluación de entonación y prosodia con las herramientas seleccionadas.
5	Comparación de los resultados entre modelos evaluados.

- OE2: Desarrollar un módulo que permita realizar la conversión de texto a voz para generación de diálogos empleando entonación y prosodia para el asistente robótico.

No.	**Actividad**
1	Diseño del módulo de conversión de texto a voz (estructura, funciones, etc.).
2	Preparación y preprocesamiento del dataset (Audios en español de Latinoamérica)
3	Entrenamiento del modelo (RNN/LSTM o similar) con ajustes de prosodia y empatía
4	Integración del modelo TTS en el asistente robótico
5	Pruebas internas y ajustes del módulo TTS (ajuste de prosodia, optimización)

- OE3: Ejecutar un plan de experimentación que permita medir el nivel de percepción que tienen los usuarios en relación al módulo de conversión de texto a voz.

No.	Actividad
1	Diseño del plan de experimentación (métricas de calidad, criterios de evaluación, selección de usuarios).
2	Realización de pruebas con usuarios (escuchar muestras de TTS, encuestas de percepción).
3	Análisis de resultados y evaluación de la percepción de los usuarios.
4	Ajustes finales al modelo y optimización basada en el feedback de usuarios.
5	Documentación del proceso de experimentación y resultados finales.

Seguidamente, se presenta el cronograma respectivo para los tres objetivos específicos:

Objetivo	Actividad	Fecha Inicio	Fecha Fin	Horas por responsable	Responsables
● OE1: Estudiar y evaluar las principales redes neuronales de aprendizaje profundo y herramientas para la generación de texto a voz, considerando aspectos de entonación y prosodia, mediante la realización de pruebas comparativas.					
OE1	1. Investigación sobre redes neuronales de aprendizaje profundo (RNN, LSTM, Transformer)	23/09/2024	30/09/2024	30	Sebastián Andrade - Kaar Mashingashi
OE1	2. Análisis de modelos preentrenados para TTS (Tacotron, Coqui TTS, etc.)	01/10/2024	07/10/2024	30	Sebastián Andrade - Kaar Mashingashi
OE1	3. Pruebas iniciales con modelos pre entrenados y evaluación de compatibilidad con español de Latinoamérica.	08/10/2024	13/10/2024	30	Sebastián Andrade - Kaar Mashingashi
OE1	4. Evaluación de entonación y prosodia con las herramientas seleccionadas	14/10/2024	18/10/2024	20	Sebastián Andrade - Kaar Mashingashi
OE1	5. Comparación de los resultados entre modelos evaluados	19/10/2024	24/10/2024	20	Sebastián Andrade - Kaar Mashingashi
● OE2: Desarrollar un módulo que permita realizar la conversión de texto a voz para generación de diálogos empleando entonación y prosodia para el asistente robótico.					
OE2	1. Planificación del módulo de conversión de texto a voz. (estructura, funciones, etc.)	25/10/2024	01/11/2024	30	Sebastián Andrade - Kaar Mashingashi
OE2	2. Preparación y preprocesamiento del dataset (audios en español de Latinoamérica)	02/11/2024	13/11/2024	40	Sebastián Andrade - Kaar Mashingashi
OE2	3. Entrenamiento del modelo (RNN/LSTM o similar) con ajustes de prosodia y empatía	14/11/2024	28/11/2024	70	Sebastián Andrade - Kaar Mashingashi
OE2	4. Preparación de estándares para el modelo TTS en el asistente robótico.	29/11/2024	05/12/2024	40	Sebastián Andrade - Kaar Mashingashi
OE2	5. Pruebas internas y ajustes	06/11/2024	20/12/2024	30	Sebastián

	del módulo TTS (ajuste de prosodia, optimización)				Andrade - Kaar Mashingashi
● OE3: Ejecutar un plan de experimentación que permita medir el nivel de percepción que tienen los usuarios en relación al módulo de conversión de texto a voz.					
OE3	1. Diseño del plan de experimentación (métricas de calidad, criterios de evaluación, selección de usuarios)	21/12/2024	29/12/2024	20	Sebastián Andrade - Kaar Mashingashi
OE3	2. Realización de pruebas con usuarios (escuchar muestras de TTS, encuestas de percepción)	30/12/2024	15/01/2025	40	Sebastián Andrade - Kaar Mashingashi
OE3	3. Análisis de resultados y evaluación de la percepción de los usuarios	16/01/2025	19/01/2025	30	Sebastián Andrade - Kaar Mashingashi
OE3	4. Ajustes finales al modelo y optimización basada en el feedback de usuarios	20/01/2025	27/01/2025	30	Sebastián Andrade - Kaar Mashingashi
OE3	5. Documentación del proceso de experimentación y resultados finales	28/01/2025	04/02/2025	20	Sebastián Andrade - Kaar Mashingashi
HORAS TOTALES				**480**	**Sebastián Andrade - Kaar Mashingashi**

VIII. Presupuesto.

DENOMINACIÓN	CANT.	COSTO UNITARIO	COSTO TOTAL
1. Bienes			
Micrófono de Condensador (Calidad para grabación	1	$150.00	$150.00
Audífonos profesionales (Monitoreo de grabación)	1	$100.00	$100.00
Almacenamiento externo (Disco duro externo, 1TB)	1	$80.00	$80.00
Subtotal Bienes			**$330.00**
2. Tecnológico			
Computadora portátil (mínimo Core i7, 16GB RAM, GPU)	1	$1,200.00	$1,200.00
Subscripción a servicios en la nube para procesamiento avanzado (Colab Pro Plus, 4 meses)	1	$50.00	$200.00
Unidades de procesamiento adicionales (para cubrir límites en servicios en la nube)	4	$50.00	$200.00
Subtotal Tecnológico			**$1,600.00**
3. Servicios			
Capacitación externa (Cursos en línea sobre redes neuronales y TTS)	1	$200.00	$200.00
Transporte para desplazamiento a reuniones y pruebas (4 meses)	4 meses	$50.00 por mes	$200.00
Subtotal Servicios			**$400.00**
4. Personal			
Trabajo como programador de proyecto TTS (480 horas en 4 meses)	1	$20.00 por hora	$9,600.00
Adquisición de herramientas especializadas para pruebas (software adicional, licencias y configuraciones avanzadas)	50 horas	$18.00 por hora	$900.00
Subtotal Personal			**$10,500.00**
5. Otros			
Material bibliográfico (artículos científicos, libros, papers especializados)	10	$40.00	$400.00
Imprevistos (5% del total del presupuesto)	1	$880.00	$880.00
Servicios de internet	4 meses	$28.00 por mes	$112.00

Subtotal Otros	**$1,392.00**
TOTAL	**$14,222.00**

IX. Conclusiones.

El desarrollo de este proyecto ha representado un avance significativo en la implementación de un sistema de conversión de texto a voz (TTS) con características de naturalidad y empatía adaptadas al español latinoamericano. A través del diseño y entrenamiento de modelos de aprendizaje profundo, se ha logrado clonar una voz con mejoras en términos de claridad, fluidez y expresividad, factores clave para su integración en entornos educativos y de asistencia personalizada.

Uno de los logros clave fue la construcción de un dataset propio compuesto por más de 700 muestras de audio, lo que permitió capturar variaciones prosódicas y dialectales de la región. Esta base de datos facilitó el entrenamiento de modelos que reflejan con mayor precisión la

diversidad del español hablado en América Latina, abordando una limitación común en los sistemas comerciales de TTS. La metodología CRISP-DM estructuró eficientemente la recolección y análisis de datos, mientras que Scrum permitió un desarrollo iterativo y adaptable, asegurando mejoras continuas en la arquitectura del modelo.

Los resultados obtenidos en las evaluaciones cuantitativas reflejan mejoras significativas respecto a los modelos iniciales. La curva de energía del modelo final mostró una estabilidad del 85%, reduciendo en un 60% las fluctuaciones abruptas observadas en versiones previas, lo que permitió un control más preciso de la entonación y la intensidad de la voz. Además, los análisis de forma de onda evidenciaron una disminución del 40% en los cortes abruptos y errores en la segmentación del habla, incrementando la fluidez y naturalidad del audio generado. La comparación de espectrogramas reveló un aumento del 30% en la continuidad de las bandas espectrales, reflejando una mayor precisión en la síntesis de fonemas.

Desde la perspectiva de los usuarios, el 75% de los evaluadores calificaron la prosodia (entonación, acento, ritmo y pausas) del modelo final como "alta" o "muy alta" en expresividad emocional, mientras que se destacó la naturalidad del habla sintetizada como adecuada para su aplicación en entornos educativos con un 70% de aprobación. La validación de la calidad del modelo se realizó mediante un análisis de concordancia utilizando el coeficiente de Kendall ($W = 0.703$), lo que indica un acuerdo fuerte entre las evaluaciones de expertas en educación infantil, respaldando la efectividad del sistema en términos de inteligibilidad y adecuación educativa.

Si bien los avances han sido significativos, aún existen oportunidades para mejorar la modulación prosódica y la variabilidad tonal del sistema. Se recomienda continuar optimizando el modelo mediante técnicas de ajuste fino con datos específicos de entonación emocional y métodos de transferencia de estilo de habla. Además, futuras investigaciones podrían enfocarse en la personalización del TTS para distintos contextos de uso, permitiendo adaptaciones específicas para asistentes virtuales, narraciones de audiolibros o tecnologías de apoyo para personas con discapacidad visual.

Por lo tanto, este proyecto representa un avance fundamental en la evolución de los sistemas TTS en español latinoamericano, demostrando que es posible lograr una síntesis de voz más natural, expresiva y culturalmente representativa. Su impacto potencial no se limita al ámbito educativo, sino que puede extenderse a aplicaciones en accesibilidad, asistencia virtual, inteligencia artificial conversacional y tecnologías inclusivas, fomentando una mayor integración de soluciones tecnológicas en la vida cotidiana y fortaleciendo la interacción entre humanos y sistemas automatizados.

X. Recomendaciones.

Para fortalecer y ampliar los hallazgos obtenidos en este estudio, se presentan las siguientes recomendaciones, dirigidas tanto a optimizar el rendimiento del módulo de conversión de texto a voz como a guiar futuras investigaciones y aplicaciones en el campo de la síntesis de voz en español latinoamericano.

En primer lugar, se recomienda continuar con la optimización del modelo de aprendizaje profundo, enfocándose en mejorar la expresividad y variabilidad emocional de la voz sintetizada. A pesar de que el sistema actual logra una entonación natural y comprensible, se identificaron ciertas limitaciones en la modulación afectiva, lo que podría abordarse mediante el entrenamiento de redes neuronales que integren control de emociones, permitiendo generar una voz que se adapte de manera dinámica al contexto del diálogo. Para ello, se sugiere explorar arquitecturas como VITS (Variational Inference Text-to-Speech) o modelos basados en transformadores, los cuales han mostrado avances en la adaptación prosódica en otros idiomas.

Asimismo, se recomienda la ampliación del dataset utilizado para el entrenamiento del modelo. Si bien la base de datos actual permitió generar una síntesis de voz con características adecuadas para el español latinoamericano, sería beneficioso incrementar el volumen y la diversidad de las muestras de audio, incorporando registros de distintos hablantes y estilos de pronunciación. Esto no solo mejoraría la generalización del modelo, sino que también permitiría adaptar el sistema a una mayor cantidad de usuarios y escenarios de interacción.

Otro aspecto a considerar es la implementación de estrategias de optimización para reducir los tiempos de inferencia y la carga computacional del sistema. La actual implementación basada en Tacotron 2 y HiFi-GAN requiere un procesamiento considerable, lo que puede representar una limitación para su uso en dispositivos con recursos restringidos. En este sentido, se recomienda explorar técnicas como la cuantización de modelos, el uso de redes neuronales más ligeras o la integración de frameworks de aceleración como TensorRT u ONNX Runtime, lo que permitiría mejorar la eficiencia del sistema sin comprometer la calidad del audio sintetizado.

Desde el punto de vista de la evaluación, se sugiere realizar pruebas adicionales con usuarios en escenarios de aplicación reales. La validación realizada con expertos proporcionó una primera aproximación sobre la percepción de la calidad del audio generado; sin embargo, sería relevante incluir pruebas con usuarios finales en entornos educativos y de asistencia personalizada para medir el impacto del sistema en la experiencia de interacción humano-máquina. Para ello, se podrían emplear metodologías de evaluación perceptual más detalladas, como escalas MOS (Mean Opinion Score) y pruebas A/B con modelos de referencia.

Además, considerando la creciente integración de asistentes virtuales en diversas aplicaciones, se recomienda explorar la combinación del sistema de síntesis de voz con modelos avanzados de procesamiento de lenguaje natural (PLN). La incorporación de modelos como GPT o T5 permitiría dotar al asistente robótico Volk de una mayor coherencia en sus respuestas, logrando una interacción más fluida y contextualizada. Esto facilitaría la implementación del módulo en ámbitos como la educación, la atención a personas con discapacidad o la asistencia en entornos médicos.

Finalmente, para garantizar la sostenibilidad y escalabilidad del sistema, se sugiere documentar de manera detallada cada etapa del proceso de desarrollo, incluyendo la configuración del modelo, los parámetros utilizados en el entrenamiento y las estrategias de ajuste implementadas.

XI. Referencias bibliográficas.

Aquilué Rubio, R. (2024). Asistente de voz local.

Camero Amador, R. D. J., Ramos Rossetes, J. J., & Mejía Suárez, O. Á. (2021). Implementación de clonador de voz en tiempo real para la lengua española usando algoritmos de aprendizaje profundo.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Ferrero Fernández, M. (2024). Previsión de demanda eléctrica usando técnicas de Aprendizaje Automático.

Garzon Montoya, J. C., Torres Argüello, L. A., & Vásquez Puerto, C. C. (2024). Aplicación de metodologías ágiles para el incremento del valor organizacional Caso: Proescon sas.

Gómez, A., García, L., & Pérez, M. (2023). Towards empathetic interaction: The role of prosody in voice interfaces. Journal of Human-Machine Interaction, 12(2), 45-59.

Gómez, A., Ruiz, A., & Castro, J. (2023). Avances recientes en síntesis de voz y su impacto en la interacción humano-robot. Revista de Inteligencia Artificial y Robótica, 15(1), 45-60.

Gómez Convers, G. H. (2024). Estudio de la relación entre los valores sociales y la aceptación de sobornos como conducta corrupta: un estudio con modelos SEM y datos de la encuesta mundial de valores.

González, M., & López, C. (2022). The impact of dialectal diversity on speech synthesis in Latin American Spanish. Journal of Phonetics and Speech Processing, 10(1), 23-31.

González, M., & López, P. (2022). Aplicaciones de asistentes robóticos en contextos educativos y de asistencia. Tecnologías y Educación, 12(3), 112-127.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hansen, S. T. (2023). Guía para principiantes de robótica e inteligencia artificial. Saxo Publish.

Heymann, H. (2021). Despliegue de modelos de aprendizaje automático para la predicción de la calidad en la producción.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hernández, P., & Pérez, R. (2020). Percepción de la calidad en sistemas de síntesis de voz para asistentes virtuales. Revista de Interacción Humano-Computadora, 18(4), 321-335.

Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing. Prentice Hall.

Jaramillo, A., & Arias, H. P. P. (2015). Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. Revista Tecnológica-ESPOL, 28(1).

King, S., & Karaiskos, V. (2019). Emotional speech synthesis: A review and a framework for future research. Speech Communication Journal, 45(3), 85-104.

Kumar, P., & Patel, R. (2024). Dynamic Emotional Adjustment in Text-to-Speech Systems. Artificial Intelligence Review, 53(4), 1115-1130. doi:10.1007/s10462-023-10398-6

Lee, J., Kim, S., & Park, H. (2023). The Effect of Vocal Frequency on Emotional Perception in Synthetic Speech. International Journal of Human-Computer Interaction, 39(2), 85-102. doi:10.1080/10447318.2023.1102253

Li, X., Wang, Y., & Liu, Y. (2023). SpeechT5: Transformer-based Speech Processing with Pretrained Models. Proceedings of the IEEE Conference on Audio, Speech, and Signal Processing.

López, A., Fernández, P., & García, M. (2023). Avances en síntesis de voz mediante Tacotron 2 en el contexto del español latino. Revista Iberoamericana de Tecnología, 15(2), 56-67.

Maskavizan, A. J., Poco, A. N., & Calzolari, A. (2023). Aspectos prácticos del uso del coeficiente de concordancia W de Kendall para el jueceo de cuestionarios en enfermería.

Martínez, J., Rodríguez, S., & Núñez, A. (2022). Implementación de WaveNet en la síntesis de voz en tiempo real. Revista de Inteligencia Artificial, 20(3), 145-160.

Morales, E., Fernández, R., & Ortega, M. (2021). Interfaces naturales y fluidas en la inteligencia artificial: Un panorama actual. Journal of Artificial Intelligence Research, 30(2), 78-94.

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

Padilla, X. A. (2022). La voz como reacción emocional: de qué nos informa la prosodia. Spanish in Context, 19(1), 72-98.

Pérez Hernández, B. (2024). Diseño ágil de aplicaciones de realidad virtual y agentes asistidos por voz utilizando inteligencia artificial y tecnología WebXR (Master's thesis).

Sánchez Trujillo, M. G., & Pérez Hernández, J. Á. (2023). Metodología CRISP-DM en la gestión de proyecto de Data Mining. Caso enfermedades dermatológicas.

Salazar, E. Y. H., & Beltrán, C. A. (2020). SCRUM, Un enfoque práctico de metodología ágil para la ingeniería de software. Tecnología Investigación y Academia, 8(2), 61-73

Salgado García, B. (2024). Aplicaciones de la Inteligencia Artificial Generativa (IAG) en el Contexto de la Seguridad.

Shen, J., et al. (2018). Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions. arXiv preprint arXiv:1712.05884.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2011). Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. Wiley.

Smith, A., Gómez, R., & Hernández, P. (2024). Adaptive Emotional Synthesis in AI-Driven Voice Assistants. Journal of Applied AI Research, 10(1), 50-66. doi:10.5555/jaair2024.1010

Schröer, C., Kruse, F., y Gómez, JM (2021). Una revisión sistemática de la literatura sobre la aplicación del modelo de proceso CRISP-DM. Procedia Computer Science , 181 , 526-534.

Tenés Trillo, E. (2023). Sistema para generar recomendaciones de cursos en una plataforma de formación.

Torres, J. I. S., & Cardenas, E. G. (2021). Análisis y aplicación de algoritmos de minería de datos. perspectivas, 6(21), 71-88.

TTS: Deep learning for Text to Speech (Discussion forum: https://discourse.mozilla.org/c/tts). (n.d.).

UNESCO. (2024, septiembre 10). La inteligencia artificial y los futuros del aprendizaje. Unesco.org. https://dataviz.unesco.org/es/digital-education/ai-future-learning

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Proceedings of the 31st Conference on Neural Information Processing Systems, 5998-6008.

Vista de Las Tecnologías Emergentes en la Sociedad del Aprendizaje. (s/f). Edu.ec. Recuperado el 21 de enero de 2025, de https://revistas.pucese.edu.ec/hallazgos21/article/view/511/436

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on Data Mining.

XII. Anexos.

Anexo 1. Glosario

- **TTS (text-to-speech):** La metodología que convierte texto escrito en voz sintetizada. Se utiliza ampliamente en asistentes virtuales y asistentes robóticos para proporcionar interacciones más naturales entre humanos y máquinas.
- **Prosodia:** Conjunto de características suprasegmentales como la entonación, el ritmo y el acento que contribuyen a la percepción de la naturalidad y emoción del sonido.
- **RNN (Red neuronal recurrente):** Este es una red neuronal artificial que puede procesar secuencias de datos como texto o audio. Se utiliza ampliamente en tareas como síntesis y reconocimiento de voz.
- **Tacotron:** un modelo de síntesis de voz que convierte texto en espectrograma y luego en audio, lo que da como resultado un sonido más natural.
- **WaveNet:** un modelo generativo desarrollado por DeepMind que crea formas de onda de audio a partir de espectrogramas para mejorar la calidad del habla producida en los sistemas TTS.

- **Conjuntos de datos:** Los datasets se utilizan para entrenar modelos de aprendizaje profundo. En un entorno TTS, un conjunto de datos normalmente consta de grabaciones de audio y sus correspondientes transcripciones de texto.
- **CRISP-DM:** una metodología estándar para la exploración de datos, que incluye fases de comprensión empresarial, análisis y preparación de datos, modelado, evaluación e implementación.
- **Scrum:** Es un marco ágil que nos permite gestionar proyectos complejos que permite la entrega continua de valor a través de ciclos de trabajo cortos llamados sprints.
- **Volk Robot Assistant:** un robot desarrollado por la Universidad Politécnica de Salesia (UPS) durante su presidencia de la UNESCO, diseñado para interactuar con los usuarios mediante voz sintetizada.
- **Síntesis de voz:** el proceso de generar voz artificial utilizando algoritmos de aprendizaje profundo o uniendo clips de voz pregrabados.
- **Transformer:** un modelo de alto nivel para tareas de procesamiento de lenguaje natural y síntesis de voz que captura de manera eficiente dependencias remotas entre texto y audio.
- **Espectrograma Mel:** representación gráfica de las frecuencias contenidas en una señal de audio. En el modelo TTS, se utiliza para convertir texto en una representación y luego en sonido.
- **Percepción del usuario:** Percepción con y para los usuarios sobre la naturalidad y calidad del habla producida por el sistema TTS, según lo evaluado por experimentos de usuarios finales.

Anexo 2. Encuestas

Anexo 2.1. Encuesta: Ordoñez Vasquez Maria Elisa

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

Cátedra UNESCO Tecnologías de apoyo para la Inclusión Educativa

GIATA

Encuesta sobre la percepción de expertos acerca del sistema de generación de voces sintéticas amigables para asistentes robóticos

Esta encuesta tiene como objetivo evaluar la percepción de expertos respecto a la calidad y utilidad de la voz generada mediante redes neuronales para su implementación en un asistente robótico socialmente interactivo. Se busca determinar si dicha voz es adecuada para mejorar la atención y el aprendizaje de niños y niñas de 6 a 8 años.

- **Apellidos y Nombres:** Ordóñez Vasquez Maria Elisa
- **Indique su género:** ☒ Femenino ☐ Masculino ☐ Otro
- **Edad:** 44
- **Indique su lugar/es de trabajo:** Psicologa Edu. Terapeuta / Docente Universitaria UPS.
- **Indique su profesión:** Psicóloga Edu. / Docente.
- **Indique sus años de experiencia profesional:** 20.
- **Indique si es o no docente:** [X] Sí [ ] No
- **¿Tiene experiencia trabajando con niños de 6 a 8 años?** [X] Sí [ ] No
- **Indique en qué área de docencia trabaja:** Inclusión Educativa.

Instrucciones: Por favor, lea cada pregunta cuidadosamente y seleccione una sola opción que me refleje su opinión marcando con una "X".

Sección 1: Calidad de la voz generada

Evalúa aspectos como naturalidad, claridad, tono y prosodia (ritmo, entonación y énfasis) de la voz generada.

Pregunta	Muy baja	Baja	Moderada	Alta	Muy alta
1. ¿Qué tan natural percibe la pronunciación de las palabras en la voz generada?			X		

2. ¿Cómo calificaría la entonación de la voz generada para transmitir emociones positivas?			X		
3. ¿Qué tan clara considera que es la voz generada para facilitar la comprensión?			X		
4. ¿Cómo evaluaría el ritmo y la entonación de la voz generada para que suene fluida y expresiva?			X		

Sección 2: Impacto en el aprendizaje y atención de los niños

Explora el impacto de la voz generada en el aprendizaje, atención y emociones de los niños.

Pregunta	Nada efectiva	Poco efectiva	Moderadamente efectiva	Efectiva	Muy efectiva
5. ¿Qué tan efectiva considera que es la voz generada para captar la atención de los niños durante actividades?			X		
6. ¿Qué tan útil cree que sería esta voz para motivar a los niños en actividades interactivas?			X		

Pregunta	Nada amigable	Poco amigable	Neutral	Amigable	Muy amigable
7. ¿Qué tan amigable considera la voz generada para niños de 6 a 8 años?			X		

Pregunta	Nada útil	Poco útil	Regular	Útil	Muy útil

8. ¿Qué tan útil considera la voz generada para actividades recreativas como cuentos interactivos?			X		

Sección 3: Comparación con sistemas estándar de TTS (Text-to-Speech)

Compara la voz generada con sistemas estándar en términos de prosodia, naturalidad y adecuación para niños.

Pregunta	Mucho peor	Peor	Igual	Mejor	Mucho mejor
9. ¿Qué tan superior considera la calidad general de la voz generada (incluyendo prosodia, naturalidad y claridad) en comparación con sistemas TTS estándar?				X	

Pregunta	Nada adecuada	Poco adecuada	Neutral	Adecuada	Muy adecuada
10. ¿Considera que la voz generada es más adecuada para niños que las disponibles en sistemas TTS?				X	
11. ¿Qué tan adecuada es la entonación de la voz generada para captar el interés de los niños?				X	

Sección 4: Exploración complementaria

Preguntas complementarias que exploran otros posibles usos y ajustes para la voz generada.

Pregunta	Nada relevante	Poco relevante	Moderadamente relevante	Relevante	Muy relevante
12. ¿Qué tan relevante considera ajustar la prosodia para mejorar la interacción con niños de 6 a 8 años?				X	

Pregunta	Nada útil	Poco útil	Regular	Útil	Muy útil
13. ¿Qué tan útil considera que la voz generada sea adaptada para actividades de apoyo educativo infantil?				X	
14. ¿Cree que el sistema podría ser utilizado en el hogar para reforzar el aprendizaje de los niños?				X	
15. ¿Qué tan importante considera que la voz generada se use en aplicaciones tecnológicas educativas?				X	
16. ¿Qué tan adecuada sería esta voz para herramientas que promuevan la colaboración entre niños?				X	

Pregunta Abierta (opcional):

17. ¿Qué aspectos considera que podrían mejorarse en la voz generada para adaptarse mejor a la edad objetivo?

la tonalidad de la voz en cada palabra.

¡GRACIAS POR SU GENTIL COLABORACIÓN!

Anexo 2.2. Encuesta: Reyes Quezada Karina Priscila

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

Cátedra UNESCO Tecnologías de apoyo para la Inclusión Educativa

GIATA

Encuesta sobre la percepción de expertos acerca del sistema de generación de voces sintéticas amigables para asistentes robóticos

Esta encuesta tiene como objetivo evaluar la percepción de expertos respecto a la calidad y utilidad de la voz generada mediante redes neuronales para su implementación en un asistente robótico socialmente interactivo. Se busca determinar si dicha voz es adecuada para mejorar la atención y el aprendizaje de niños y niñas de 6 a 8 años.

- **Apellidos y Nombres:** Reyes Quezada Karina Priscila
- **Indique su género:** ☒ Femenino ☐ Masculino ☐ Otro
- **Edad:** 32
- **Indique su lugar/es de trabajo:** Kukuli, Unidad Educativa Paccha, U.P.S
- **Indique su profesión:** Docente
- **Indique sus años de experiencia profesional:** 9 años
- **Indique si es o no docente:** [x] Sí [ ] No
- **¿Tiene experiencia trabajando con niños de 6 a 8 años?** [x] Sí [ ] No
- **Indique en qué área de docencia trabaja:** Educación Inicial

Instrucciones: Por favor, lea cada pregunta cuidadosamente y seleccione una sola opción que mejor refleje su opinión marcando con una "X".

Sección 1: Calidad de la voz generada

Evalúa aspectos como naturalidad, claridad, tono y prosodia (ritmo, entonación y énfasis) de la voz generada.

Pregunta	Muy baja	Baja	Moderada	Alta	Muy alta
1. ¿Qué tan natural percibe la pronunciación de las palabras en la voz generada?			X		

2. ¿Cómo calificaría la entonación de la voz generada para transmitir emociones positivas?			X		
3. ¿Qué tan clara considera que es la voz generada para facilitar la comprensión?			X		
4. ¿Cómo evaluaría el ritmo y la entonación de la voz generada para que suene fluida y expresiva?		X			

Sección 2: Impacto en el aprendizaje y atención de los niños

Explora el impacto de la voz generada en el aprendizaje, atención y emociones de los niños.

Pregunta	Nada efectiva	Poco efectiva	Moderadamente efectiva	Efectiva	Muy efectiva
5. ¿Qué tan efectiva considera que es la voz generada para captar la atención de los niños durante actividades?			X		
6. ¿Qué tan útil cree que sería esta voz para motivar a los niños en actividades interactivas?			X		

Pregunta	Nada amigable	Poco amigable	Neutral	Amigable	Muy amigable
7. ¿Qué tan amigable considera la voz generada para niños de 6 a 8 años?				X	

Pregunta	Nada útil	Poco útil	Regular	Útil	Muy útil

8. ¿Qué tan útil considera la voz generada para actividades recreativas como cuentos interactivos?				X	

Sección 3: Comparación con sistemas estándar de TTS (Text-to-Speech)

Compara la voz generada con sistemas estándar en términos de prosodia, naturalidad y adecuación para niños.

Pregunta	Mucho peor	Peor	Igual	Mejor	Mucho mejor
9. ¿Qué tan superior considera la calidad general de la voz generada (incluyendo prosodia, naturalidad y claridad) en comparación con sistemas TTS estándar?					X

Pregunta	Nada adecuada	Poco adecuada	Neutral	Adecuada	Muy adecuada
10. ¿Considera que la voz generada es más adecuada para niños que las disponibles en sistemas TTS?					X
11. ¿Qué tan adecuada es la entonación de la voz generada para captar el interés de los niños?				X	

Sección 4: Exploración complementaria

Preguntas complementarias que exploran otros posibles usos y ajustes para la voz generada.

Pregunta	Nada relevante	Poco relevante	Moderadamente relevante	Relevante	Muy relevante
12. ¿Qué tan relevante considera ajustar la prosodia para mejorar la interacción con niños de 6 a 8 años?				X	

Pregunta	Nada útil	Poco útil	Regular	Útil	Muy útil
13. ¿Qué tan útil considera que la voz generada sea adaptada para actividades de apoyo educativo infantil?					X
14. ¿Cree que el sistema podría ser utilizado en el hogar para reforzar el aprendizaje de los niños?					X
15. ¿Qué tan importante considera que la voz generada se use en aplicaciones tecnológicas educativas?					X
16. ¿Qué tan adecuada sería esta voz para herramientas que promuevan la colaboración entre niños?					X

Pregunta Abierta (opcional):

17. ¿Qué aspectos considera que podrían mejorarse en la voz generada para adaptarse mejor a la edad objetivo?

Más pausa y mayor entonación

¡GRACIAS POR SU GENTIL COLABORACIÓN!

Anexo 2.3. Encuesta: Almeida Soliz Nidia Elizabeth

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

Cátedra UNESCO Tecnologías de apoyo para la Inclusión Educativa

GIATA

Encuesta sobre la percepción de expertos acerca del sistema de generación de voces sintéticas amigables para asistentes robóticos

Esta encuesta tiene como objetivo evaluar la percepción de expertos respecto a la calidad y utilidad de la voz generada mediante redes neuronales para su implementación en un asistente robótico socialmente interactivo. Se busca determinar si dicha voz es adecuada para mejorar la atención y el aprendizaje de niños y niñas de 6 a 8 años.

- **Apellidos y Nombres:** Almeida Soliz Nidia Elizabeth
- **Indique su género:** ☑ Femenino ☐ Masculino ☐ Otro
- **Edad:** 44 años
- **Indique su lugar/es de trabajo:** U.E. Eugenio Espejo y UPS.
- **Indique su profesión:** Lic. en Ciencias de la Educación, mención Educación Infantil
- **Indique sus años de experiencia profesional:** 21 años.
- **Indique si es o no docente:** [✓] Sí [ ] No
- **¿Tiene experiencia trabajando con niños de 6 a 8 años?** [✓] Sí [ ] No
- **Indique en qué área de docencia trabaja:** Educación Inicial y Preparatoria

Instrucciones: Por favor, lea cada pregunta cuidadosamente y seleccione una sola opción que me refleje su opinión marcando con una "X".

Sección 1: Calidad de la voz generada

Evalúa aspectos como naturalidad, claridad, tono y prosodia (ritmo, entonación y énfasis) de la voz generada.

Pregunta	Muy baja	Baja	Moderada	Alta	Muy alta
1. ¿Qué tan natural percibe la pronunciación de las palabras en la voz generada?			X		

2. ¿Cómo calificaría la entonación de la voz generada para transmitir emociones positivas?			X		
3. ¿Qué tan clara considera que es la voz generada para facilitar la comprensión?			X		
4. ¿Cómo evaluaría el ritmo y la entonación de la voz generada para que suene fluida y expresiva?			X		

Sección 2: Impacto en el aprendizaje y atención de los niños

Explora el impacto de la voz generada en el aprendizaje, atención y emociones de los niños.

Pregunta	Nada efectiva	Poco efectiva	Moderadamente efectiva	Efectiva	Muy efectiva
5. ¿Qué tan efectiva considera que es la voz generada para captar la atención de los niños durante actividades?			X		
6. ¿Qué tan útil cree que sería esta voz para motivar a los niños en actividades interactivas?					✓

Pregunta	Nada amigable	Poco amigable	Neutral	Amigable	Muy amigable
7. ¿Qué tan amigable considera la voz generada para niños de 6 a 8 años?				✓	

Pregunta	Nada útil	Poco útil	Regular	Útil	Muy útil

8. ¿Qué tan útil considera la voz generada para actividades recreativas como cuentos interactivos?					X

Sección 3: Comparación con sistemas estándar de TTS (Text-to-Speech)

Compara la voz generada con sistemas estándar en términos de prosodia, naturalidad y adecuación para niños.

Pregunta	Mucho peor	Peor	Igual	Mejor	Mucho mejor
9. ¿Qué tan superior considera la calidad general de la voz generada (incluyendo prosodia, naturalidad y claridad) en comparación con sistemas TTS estándar?				X	

Pregunta	Nada adecuada	Poco adecuada	Neutral	Adecuada	Muy adecuada
10. ¿Considera que la voz generada es más adecuada para niños que las disponibles en sistemas TTS?					X
11. ¿Qué tan adecuada es la entonación de la voz generada para captar el interés de los niños?					X

Sección 4: Exploración complementaria

Preguntas complementarias que exploran otros posibles usos y ajustes para la voz generada.

Pregunta	Nada relevante	Poco relevante	Moderadamente relevante	Relevante	Muy relevante
12. ¿Qué tan relevante considera ajustar la prosodia para mejorar la interacción con niños de 6 a 8 años?					X

Pregunta	Nada útil	Poco útil	Regular	Útil	Muy útil
13. ¿Qué tan útil considera que la voz generada sea adaptada para actividades de apoyo educativo infantil?					X
14. ¿Cree que el sistema podría ser utilizado en el hogar para reforzar el aprendizaje de los niños?					X
15. ¿Qué tan importante considera que la voz generada se use en aplicaciones tecnológicas educativas?					X
16. ¿Qué tan adecuada sería esta voz para herramientas que promuevan la colaboración entre niños?					X

Pregunta Abierta (opcional):

17. ¿Qué aspectos considera que podrían mejorarse en la voz generada para adaptarse mejor a la edad objetivo?

Tener un ritmo que varíe en función de cada temática.

La entonación de palabras.

¡GRACIAS POR SU GENTIL COLABORACIÓN!

Anexo 2.4. Encuesta: Zamora Torres Johana Elizabeth}

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

Cátedra UNESCO Tecnologías de apoyo para la Inclusión Educativa

GIATA

Encuesta sobre la percepción de expertos acerca del sistema de generación de voces sintéticas amigables para asistentes robóticos

Esta encuesta tiene como objetivo evaluar la percepción de expertos respecto a la calidad y utilidad de la voz generada mediante redes neuronales para su implementación en un asistente robótico socialmente interactivo. Se busca determinar si dicha voz es adecuada para mejorar la atención y el aprendizaje de niños y niñas de 6 a 8 años.

- **Apellidos y Nombres:** Zamora Torres Johana Elizabeth
- **Indique su género:** ☑ Femenino ☐ Masculino ☐ Otro
- **Edad:** 44
- **Indique su lugar/es de trabajo:** UPS Carrera de Educación
- **Indique su profesión:** Mgst. Intervención y Educación Inic
- **Indique sus años de experiencia profesional:** 30
- **Indique si es o no docente:** [✓] Sí [ ] No
- **¿Tiene experiencia trabajando con niños de 6 a 8 años?** [✓] Sí [ ] No
- **Indique en qué área de docencia trabaja:** Educación Inicial y Universitari

Instrucciones: Por favor, lea cada pregunta cuidadosamente y seleccione una sola opción que me refleje su opinión marcando con una "X".

Sección 1: Calidad de la voz generada

Evalúa aspectos como naturalidad, claridad, tono y prosodia (ritmo, entonación y énfasis) de la voz generada.

Pregunta	Muy baja	Baja	Moderada	Alta	Muy alta
1. ¿Qué tan natural percibe la pronunciación de las palabras en la voz generada?		X			

2. ¿Cómo calificaría la entonación de la voz generada para transmitir emociones positivas?		X			
3. ¿Qué tan clara considera que es la voz generada para facilitar la comprensión?		X			
4. ¿Cómo evaluaría el ritmo y la entonación de la voz generada para que suene fluida y expresiva?		X			

Sección 2: Impacto en el aprendizaje y atención de los niños

Explora el impacto de la voz generada en el aprendizaje, atención y emociones de los niños.

Pregunta	Nada efectiva	Poco efectiva	Moderadamente efectiva	Efectiva	Muy efectiva
5. ¿Qué tan efectiva considera que es la voz generada para captar la atención de los niños durante actividades?		X			
6. ¿Qué tan útil cree que sería esta voz para motivar a los niños en actividades interactivas?		X			

Pregunta	Nada amigable	Poco amigable	Neutral	Amigable	Muy amigable
7. ¿Qué tan amigable considera la voz generada para niños de 6 a 8 años?		X			

Pregunta	Nada útil	Poco útil	Regular	Útil	Muy útil

8. ¿Qué tan útil considera la voz generada para actividades recreativas como cuentos interactivos?		X			

Sección 3: Comparación con sistemas estándar de TTS (Text-to-Speech)

Compara la voz generada con sistemas estándar en términos de prosodia, naturalidad y adecuación para niños.

Pregunta	Mucho peor	Peor	Igual	Mejor	Mucho mejor
9. ¿Qué tan superior considera la calidad general de la voz generada (incluyendo prosodia, naturalidad y claridad) en comparación con sistemas TTS estándar?				✓	

Pregunta	Nada adecuada	Poco adecuada	Neutral	Adecuada	Muy adecuada
10. ¿Considera que la voz generada es más adecuada para niños que las disponibles en sistemas TTS?			✓		
11. ¿Qué tan adecuada es la entonación de la voz generada para captar el interés de los niños?		✓			

Sección 4: Exploración complementaria

Preguntas complementarias que exploran otros posibles usos y ajustes para la voz generada.

Pregunta	Nada relevante	Poco relevante	Moderadamente relevante	Relevante	Muy relevante
12. ¿Qué tan relevante considera ajustar la prosodia para mejorar la interacción con niños de 6 a 8 años?		✓			

Pregunta	Nada útil	Poco útil	Regular	Útil	Muy útil
13. ¿Qué tan útil considera que la voz generada sea adaptada para actividades de apoyo educativo infantil?		✓			
14. ¿Cree que el sistema podría ser utilizado en el hogar para reforzar el aprendizaje de los niños?				✓	
15. ¿Qué tan importante considera que la voz generada se use en aplicaciones tecnológicas educativas?		✓			
16. ¿Qué tan adecuada sería esta voz para herramientas que promuevan la colaboración entre niños?		✓			

Pregunta Abierta (opcional):

17. ¿Qué aspectos considera que podrían mejorarse en la voz generada para adaptarse mejor a la edad objetivo?

La voz debe ser más natural, que llame la atención de los niños, aún los pequeños y es necesario moverlos.

¡GRACIAS POR SU GENTIL COLABORACIÓN!

Anexo 3. Soporte Fotográfico

Fotografía 1. Exposición del Proyecto

Fotografía 2. Desarrollo de la Encuesta

Fotografía 3. Retroalimentación del proyecto por la docente