



Facultad de Ciencias de la Administración

**Carrera de Ingeniería en Ciencias de la
Computación**

**AJUSTE FINO DE UN MODELO PRE-
ENTRENADO PARA RECONOCIMIENTO DE
VOZ INFANTIL CON VALIDACIÓN SEMÁNTICA
BASADA EN MODELOS DE LENGUAJE**

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

Autor:

James Rusbel García Peñaloza

Director:

Fabián Marcelo Carvajal Vargas

Cuenca – Ecuador

2025

DEDICATORIA

Dedico este trabajo de titulación a mis padres, por su guía, apoyo incondicional y ejemplo de esfuerzo que me han inspirado a alcanzar mis objetivos académicos y personales.

A mi familia, por su comprensión y motivación constante durante el desarrollo de esta investigación.

Extiendo también esta dedicatoria a todas las personas que, de una u otra manera, contribuyeron a la culminación de este proyecto, brindando su tiempo, conocimiento y aliento en los momentos más exigentes del proceso.

AGRADECIMIENTO

Expreso mi más sincero agradecimiento a Dios, por brindarme la fortaleza, la sabiduría y la perseverancia necesarias para culminar con éxito esta etapa académica.

A mis padres y familia, por su apoyo incondicional, su comprensión y su constante motivación a lo largo de todo este proceso.

Extiendo mi más profundo reconocimiento al Ing. Fabián Marcelo Carvajal Vargas, director de tesis, por su guía, paciencia y valiosas orientaciones durante el desarrollo de esta investigación.

De igual manera, manifiesto mi gratitud al Ing. Patricio Santiago García Montero y a la Ing. María Inés Acosta Urigüen, por su colaboración, asesoramiento técnico y disposición para compartir sus conocimientos, los cuales fueron fundamentales para el avance y culminación de este trabajo.

Finalmente, agradezco a mis docentes, compañeros y amigos, por su apoyo y compañía en esta importante etapa de mi formación profesional.

Índice de Contenidos

DEDICATORIA	i
AGRADECIMIENTO	ii
Índice de Contenidos	iii
Índice de Figuras	v
Índice de Tablas	vi
Índice de Anexos	vii
RESUMEN	viii
ABSTRACT	viii
1. Introducción	1
1.1 Objetivos	1
1.1.1 Objetivo General	1
1.1.2 Objetivos Específicos	1
2. Marco Teórico y Estado del Arte	1
2.1 Marco Teórico	1
2.1.1 Introducción a los Datos de Audio	2
2.1.2 Muestreo y Frecuencia de Muestreo	2
2.1.3 Amplitud y Profundidad de Bits	2
2.1.4. Espectro de Frecuencia	3
2.1.5. Espectrograma	3
2.1.6. Espectrograma de Mel	3
2.1.7. Remuestreo de los Datos de Audio	4
2.1.8. Funciones de Audio en el Dominio del Tiempo	4
2.1.9. Filtrado del Conjunto de Datos	4
2.1.10. Aumentación de Datos	5
2.1.11. Aplicaciones de Audio: Reconocimiento Automático de Voz, Diarización y Otras Tareas Relacionadas	5
2.1.12. Arquitectura de Transformers para Audio	5
2.1.13. Arquitectura CTC	6
2.1.14. Arquitectura Seq2Seq	6
2.1.15. Modelos Previamente Entrenados para ASR	7
2.1.16. Elección de un Conjunto de Datos	7
2.1.17. Evaluación y Métricas para el Reconocimiento de Voz	8
2.1.18. Reprocesamiento de un Conjunto de Datos de Audio	8
2.1.19. Ajuste Fino de un Sistema ASR	9
2.2 Estado del Arte	9
3. Metodología	9
3.1 Comprensión del Problema	10
3.1.1 Análisis Pedagógico y Fundamentos Educativos	11
3.1.2 Contexto y Justificación	11
3.1.3 Actividades Realizadas	12
3.1.4 Consideraciones Éticas	12
3.2 Comprensión de Datos	12
3.2.1 Contexto y Fuentes de Datos	13
3.2.2 Estructura del Corpus y Objetivos Pedagógicos	14
3.2.3 Características del Corpus Recolectado	14
3.2.4 Procesamiento Inicial y Diarización	15

3.2.5 Transcripción y Validación de Calidad	16
3.2.6 Etiquetado de Roles y Eventos Acústicos	16
3.2.7 Aspectos Éticos y Resguardo de los Datos	17
3.3 Preparación de Datos.....	17
3.3.1 Actividades Desarrolladas.....	17
3.4 Modelado.....	19
3.4.1 Configuración del Entorno de Entrenamiento	20
3.4.2 Integración del Dataset y Preprocesamiento	20
3.4.3 Ajuste Fino (<i>fine-tuning</i>) del Modelo.....	21
3.4.4 Justificación Metodológica	21
3.4.5 Evaluación Interna del Modelo	22
3.4.6 Publicación y Respaldo del Modelo.....	22
3.5 Evaluación y Validación Semántica.....	23
3.5.1 Objetivos de la Evaluación.....	24
3.5.2 Procedimiento de Evaluación Técnica	24
3.5.3 Validación Semántica con LLM.....	25
3.6 Despliegue.....	25
3.6.1 Publicación del Modelo en Hugging Face	26
3.6.2 Integración con la Validación Semántica.....	26
3.6.3 Aplicación Pedagógica y Uso Potencial.....	26
3.6.4 Limitaciones Observadas	27
4. Resultados	27
4.1 Resultados del Entrenamiento del Modelo.....	28
4.2 Entrenamiento del Modelo	28
4.2.1 Desempeño del Modelo en la Reducción del Error de Palabra (WER).....	30
4.2.2 Comportamiento de la Función de Pérdida y Estabilidad del Aprendizaje del Modelo..	31
4.3 Uso Experimental.....	32
4.4 Evaluación y Validación Semántica.....	33
4.4.1 Justificación de la Transcripción y del Proceso	34
4.4.2 Interpretación de las Métricas	35
5. Discusión.....	36
5.1 Interpretación General de los Resultados.....	36
5.2 Análisis de los Aportes Técnicos	37
5.3 Implicaciones Pedagógicas y Aplicativas	38
5.4 Limitaciones del Estudio.....	38
6. Conclusión.....	39
7. Referencias.....	40
8. Anexos.....	49

Índice de Figuras

Figura 1 Flujo metodológico para el ajuste fino de un modelo de reconocimiento automático de voz infantil (ASR) y su validación semántica.....	10
Figura 2 Variación del Word Error Rate (WER) a lo largo del fine-tuning del modelo Whisper ..	30
Figura 3 Evolución de la pérdida de entrenamiento (Training Loss) durante el ajuste fino del modelo	32

Índice de Tablas

Tabla 1	Descripción de los campos de la base de datos unificada de clips de audio	18
Tabla 2	Categorías semánticas y descripción del etiquetado de los fragmentos de audio	19
Tabla 3	Configuración de parámetros utilizados durante el ajuste fino del modelo Whisper	21
Tabla 4	Métricas de evaluación y fórmulas empleadas en el modelo de reconocimiento automático del habla	24
Tabla 5	Resultados del entrenamiento del modelo durante las épocas.....	29
Tabla 6	Resultados de las métricas semánticas obtenidas en la evaluación del modelo	33
Tabla 7	Evaluación semántica del modelo basada en coherencia y fluidez de la transcripción.....	34

Índice de Anexos

Anexo 1 Carta de autorización institucional para la recolección de datos en el CEIAP	49
Anexo 2 Carta de autorización institucional para la recolección de datos en la Unidad Educativa Particular Estrellitas Creativas	50

Ajuste Fino De Un Modelo Pre-Entrenado Para Reconocimiento De Voz Infantil Con Validación Semántica Basada En Modelos De Lenguaje

RESUMEN

La investigación tuvo como objetivo optimizar un modelo de reconocimiento automático de voz (ASR) adaptado al habla infantil en español mediante el ajuste fino del modelo pre-entrenado Whisper-Large-V3-Turbo. El estudio se fundamentó en la corriente teórica del aprendizaje profundo y el procesamiento del lenguaje natural, integrando una validación semántica basada en modelos de lenguaje (LLM) mediante Meta/Llama 3.3. Se empleó un corpus de 299 clips de voz infantil, segmentados y anotados para entrenamiento en Google Colab con la biblioteca datasets de Hugging Face. Los resultados mostraron una reducción del Word Error Rate (WER) hasta un 81.74 % y valores semánticos adecuados en tono (0.8) y toxicidad (0.0). Se concluyó que la combinación del ajuste fino y la validación semántica permitió mejorar la comprensión contextual del habla infantil, aportando una metodología innovadora para el desarrollo de sistemas ASR con aplicaciones educativas y terapéuticas.

Palabras clave: ajuste fino, aprendizaje profundo, habla infantil, modelos de lenguaje, reconocimiento automático de voz, validación semántica, whisper.

Fine-Tuning a Pre-Trained Model for Child Voice Recognition with Semantic Validation Based on Language Models

ABSTRACT

The research aimed to optimize an automatic speech recognition (ASR) model adapted to children's speech in Spanish by fine-tuning the pre-trained Whisper-Large-V3-Turbo model. The study was based on the theoretical current of deep learning and natural language processing, integrating semantic validation based on language models (LLM) using Meta/Llama 3.3. A corpus of 299 children's voice clips, segmented and annotated for training in Google Colab with the Hugging Face datasets library, was used. The results showed a reduction in the Word Error Rate (WER) to 81.74% and adequate semantic values in tone (0.8) and toxicity (0.0). It was concluded that the combination of fine-tuning and semantic validation improved the contextual understanding of children's speech, providing an innovative methodology for the development of ASR systems with educational and therapeutic applications.

Keywords: fine-tuning, deep learning, child speech, language models, automatic speech recognition, semantic validation, whisper.