Universidad CENFOTEC

Escuela de Sistemas Inteligentes

Maestría en Ingeniería del Software con énfasis en Inteligencia Artificial

Documento final de Proyecto de Investigación Aplicada 2

Creación de un sistema de procesamiento lenguaje natural para la preservación e identificación del lenguaje Bribri

Gianfranco Bagnarello Hernández

Diciembre 2025

Tabla de Contenido

Contenidos

Frases Clave: *Bribri, idioma indígena, preservación lingüística, digitalización, procesamiento de lenguaje natural, natural language processing, NLP, Costa Rica, Talamanca.*

Capítulo 1. Introducción

1.1 Generalidades

Las lenguas indígenas están desapareciendo a un ritmo alarmante en el siglo XXI, mientras que la tecnología avanza constantemente en el campo del procesamiento de lenguaje natural. Este desarrollo tecnológico ha permitido a investigadores públicos y privados crear herramientas de transcripción automática utilizando inteligencia artificial. Sin embargo, la limitación principal para lenguas como el bribri es la escasez de audio disponible, ya que los corpus son muy pequeños, clasificándolas como "lenguajes de bajo recurso" o "*low resource languages*" debido a la falta de grandes volúmenes de datos de hablantes nativos.

Esta investigación propone utilizar el modelo pre entrenado *Whisper* de la empresa OpenAI, agregando inicialmente únicamente 10 minutos de audio bribri para lograr una transcripción exitosa, superando así las limitaciones de investigaciones previas que requieren más de una hora de audio. Este enfoque puede representar mayor eficiencia, menor poder computacional requerido, menos trabajo de campo para obtener audios, y los hallazgos pueden ser aplicables a otros lenguajes clasificados como de bajo recurso, contribuyendo así a la preservación digital de lenguas indígenas en peligro de extinción.

1.2 Antecedentes del Problema

A pesar de los avances en el procesamiento de lenguaje natural, el reconocimiento automático de voz para lenguas indígenas de bajo recurso sigue siendo un desafío significativo. Aunque existen investigaciones sobre modelos pre-entrenados como wav2vec para lenguas de bajo recurso en general, el único estudio específico de ASR para el idioma bribri fue realizado por Coto-Solano (2021), quien utilizó métodos tradicionales GMM/HMM y CTC con representaciones diseñadas manualmente, entrenando con 68 minutos de audio. Sus resultados mostraron un error de caracteres (CER) del 33% y un error de palabras (WER) del 50%, lo que equivale a que el sistema falle en la mitad de las transcripciones, siendo prácticamente inútil para aplicaciones reales.

Las limitaciones de este enfoque radican en el uso de métodos tradicionales que parten desde cero y requieren diseño manual de características específicas del idioma. Esto ha resultado en una falta de sistemas eficientes y precisos para el bribri, limitando las posibilidades de documentación y revitalización de la lengua. La ausencia de investigaciones adicionales en este campo representa un vacío crítico que afecta tanto a la comunidad bribri como al desarrollo de tecnologías inclusivas para lenguas indígenas de las Américas.

1.3 Definición y Descripción del Problema

El problema central de este proyecto surge en un contexto donde las estimaciones sugieren que entre el 50% y 95% de los idiomas hablados actualmente podrían estar extintos o en grave peligro para el año 2100, siendo las lenguas indígenas las más vulnerables (*Indigenous Children's Education and Indigenous Languages*). En este escenario crítico, existe una falta de sistemas de reconocimiento automático de voz específicamente diseñados para el idioma Bribri que puedan funcionar eficientemente con recursos de datos extremadamente limitados. Actualmente, solo existe un corpus bribri pan-dialectal con aproximadamente 68 minutos de audio total, y el único estudio específico de ASR para bribri (Coto-Solano, 2021) logró únicamente un 50% de precisión. Aunque investigaciones recientes como las del *AmericasNLP Competition (Ebrahimi et al., 2022)*

han incluido al Bribri junto con otras lenguas indígenas, no existen sistemas especializados diseñados exclusivamente para las características lingüísticas del Bribri.

Sin una solución adecuada, la comunidad Bribri permanece marginada del desarrollo tecnológico actual, perdiendo oportunidades de preservación digital y revitalización lingüística. La ausencia de herramientas de transcripción automática eficientes limita severamente las posibilidades de documentación, educación y transmisión del idioma. Esto es particularmente crítico considerando que el trabajo de campo para recolectar grandes volúmenes de datos implica costos económicos significativos y recursos humanos especializados que actualmente son escasos, perpetuando así el ciclo de invisibilidad científica y tecnológica de las lenguas indígenas.

1.4 Justificación

La justificación de este proyecto radica en la necesidad de desarrollar un sistema de procesamiento de lenguaje natural que permita la transcripción automática de audio a texto del idioma Bribri, lengua indígena que requiere herramientas tecnológicas para su preservación y documentación digital. Según C.K. Galla, en su artículo "Indigenous language revitalization, promotion, and education: function of digital technology", "En este mundo culturalmente diverso, es difícil imaginar la supervivencia de las lenguas indígenas en el siglo XXI sin la intervención de la tecnología digital. La urgencia persiste, ya que con cada lengua que desaparece, se pierde una riqueza de conocimiento que puede ser irrecuperable."

Actualmente, la ausencia de herramientas tecnológicas para la transcripción automática del Bribri genera dificultades significativas para la documentación y preservación de este patrimonio lingüístico. La falta de sistemas de PLN específicos para el Bribri limita la capacidad de crear archivos digitales textuales de manera eficiente, lo cual resulta en procesos manuales extensos y costosos para la transcripción de grabaciones, y retrasa los esfuerzos de preservación cultural mientras el idioma continúa en riesgo de extinción.

En otros países se han logrado iniciativas exitosas en la preservación digital de lenguas indígenas, como el proyecto Arctic Megapedia, donde "los pueblos indígenas del Ártico han podido participar activamente en el intercambio de información y presentar su cultura y lengua tanto en sus idiomas nativos como en lenguas de comunicación internacional. La base de datos acumulada en este portal se ha convertido en la base para el estudio de la historia, cultura y lenguas de todos los pueblos indígenas del Ártico, permitiendo la restauración de conexiones lingüísticas y culturales perdidas, así como el fortalecimiento de la identidad étnica" (A. V. Zhozhikov, "Digitalization of the cultural heritage of the indigenous peoples of the Arctic").

Al desarrollar un sistema funcional de transcripción automática para el Bribri, se espera reducir significativamente los costos económicos y temporales asociados a la transcripción manual de grabaciones, facilitar la creación de archivos digitales textuales de manera más eficiente y accesible, y proporcionar una herramienta tecnológica que apoye directamente los esfuerzos de preservación y documentación del patrimonio lingüístico Bribri. Además, se busca establecer un precedente tecnológico que pueda replicarse y adaptarse para otras lenguas indígenas de Costa Rica y Centroamérica, contribuyendo así a la digitalización más amplia del patrimonio lingüístico indígena y proporcionando una solución práctica que facilite el trabajo de investigadores, educadores y comunidades en la preservación del Bribri en formato digital.

En este caso, la investigación será de tipo aplicada, debido a que para solucionar el problema se busca desarrollar e implementar un sistema de procesamiento de lenguaje natural que logre la transcripción y reconocimiento efectivo del lenguaje Bribri en forma de audio, creando una herramienta tecnológica funcional que sea viable y útil para la comunidad.

1.5 Viabilidad

1.5.1 Punto de Vista Técnico

Desde un punto de vista técnico, el proyecto es viable mediante el uso de wav2vec 2.0 como modelo base para el reconocimiento de voz, complementado con una arquitectura que incluye un decoder entrenado en audio y texto para evaluar los resultados del modelo y realizar el refinamiento de hiper parámetros. El modelo wav2vec 2.0 fue presentado por Baevski, Zhou, Mohamed y Auli (2020) y entrenado originalmente con múltiples GPUs de alto rendimiento (por ejemplo, V100), debido a su gran complejidad y el tamaño de lote requerido para pre-entrenamiento y fine-tuning. No obstante, la arquitectura moderna Ada Lovelace, implementada en la NVIDIA RTX 4070, incorpora Tensor Cores de cuarta generación que soportan cálculo en precisión mixta (FP16, BF16) y FP8, lo que permite acelerar operaciones clave de aprendizaje profundo sin pérdida de precisión significativa. Asimismo, investigaciones de optimización como FlashAttention-2 (Dao, 2023) demuestran mejoras de hasta 2 × en velocidad y uso de FLOPs (*Floating Point Operations*) en hardware reciente, aprovechando la jerarquía de memoria y cálculos en FP16/BF16. En conjunto con un procesador Intel Core i9 de última generación, el sistema ofrece la capacidad técnica necesaria para ejecutar tareas de entrenamiento e investigación con wav2vec 2.0 a nivel académico.

La viabilidad técnica se fundamenta también en la comprobada efectividad de wav2vec 2.0 para lenguas con pocos recursos. Según Yi, Wang, Cheng, Zhou y Xu en su artículo "Applying wav2vec2.0 to speech recognition in various low-resource languages", se logran mejoras relativas de más del 20% en seis idiomas comparado con trabajos previos, donde el inglés alcanza una ganancia del 52.4%. Los autores concluyen que "wav2vec 2.0 can dynamically merge the fine-grained presentation into coarser-grained presentation to fit the target task", lo cual es especialmente relevante para lenguas indígenas como el Bribri que requieren adaptaciones específicas.

Adicionalmente, según Deshmukh, P., Kulkarni, N. et al. en su artículo "Leveraging Parameter Efficient Training Methods for Low Resource Text Classification: A case study in Marathi", "Además, utilizando estos modelos, el artículo presenta e ilustra el uso de técnicas de Ajuste Fino Eficiente en Parámetros (PEFT). Esto resuelve el importante problema de los costos computacionales en el desarrollo de modelos y acelera enormemente la velocidad de entrenamiento sin sacrificar la precisión." Estos métodos PEFT hacen viable el uso de procesamiento de lenguaje natural usando pocos recursos computacionales, lo que fortalece la viabilidad técnica del proyecto.

1.5.2 Punto de Vista Operativo

Operativamente, el proyecto es viable dado que se cuenta con acceso al Corpus pandialectal oral de la lengua bribri (bribri.net), desarrollado por Flores Solórzano (2017), que contiene aproximadamente 1 hora de audio de habla espontánea en los tres dialectos reconocidos del Bribri, junto con sus respectivas transcripciones. Este corpus documenta por primera vez el habla de la interacción cotidiana, incluyendo conversaciones, monólogos y narraciones tradicionales, proporcionando la diversidad lingüística necesaria para el entrenamiento del modelo. Los datos se encuentran disponibles en formato digital y pueden requerir preprocesamiento mínimo para su adaptación a los requerimientos del modelo wav2vec 2.0.

1.5.3 Punto de Vista Económico

Económicamente, el proyecto es viable debido a la disponibilidad de recursos existentes. Se cuenta con el hardware necesario para el entrenamiento (procesador Intel Core i9-14900HX y GPU NVIDIA RTX 4070), eliminando la necesidad de inversión en infraestructura computacional o servicios en la nube. El corpus de datos está disponible de forma gratuita para fines académicos, y las herramientas de software requeridas, incluyendo las bibliotecas de *PyTorch* y *Hugging Face Transformers*, son de código abierto. Los únicos costos asociados corresponden al tiempo de investigación y desarrollo, así como el consumo eléctrico durante el entrenamiento, los cuales son mínimos comparados con los beneficios esperados de establecer una metodología replicable para la preservación digital de lenguas indígenas con recursos limitados.

1.6 Objetivos

1.6.1 Objetivo General.

Entrenar un modelo de reconocimiento automático del habla (ASR) para el lenguaje Bribri para su transcripción.

1.6.2 Objetivos Específicos.

1. Recopilar y limpiar un conjunto de datos en idioma Bribri para su uso en el entrenamiento del sistema de transcripción.

2. Analizar aspectos específicos del Bribri que impacten el entrenamiento de wav2vec 2.0

3. Estudiar y seleccionar las técnicas de NLP más adecuadas para la transcripción automática del Bribri a texto.

4. Implementar y entrenar varios modelos de NLP utilizando los datos recopilados, con el objetivo de optimizar la transcripción.

5. Evaluar el desempeño del sistema utilizando métricas como el *Word Error Rate* (WER) y *Character Error Rat*e (CER).

6. Desarrollar un prototipo funcional del sistema de transcripción, ajustándolo según los resultados obtenidos en la evaluación.

1.7 Alcances y Limitaciones

1.7.1 Alcances

Este proyecto se centra en el desarrollo de un sistema de procesamiento de lenguaje natural específicamente para la transcripción automática de audio a texto del idioma Bribri. Se implementará utilizando modelos de reconocimiento automático de voz basados en wav2vec 2.0, adaptados y entrenados con datos disponibles del idioma Bribri. El sistema se desarrollará como una prueba de concepto funcional que demuestre la viabilidad técnica de la transcripción automática para esta lengua indígena, utilizando técnicas de refinamiento de hiper parámetros para optimizar el rendimiento con los datos disponibles.

1.7.2 Limitaciones

El proyecto trabajará con un conjunto limitado de datos de audio en Bribri obtenidos de fuentes públicas y disponibles en internet, lo cual puede afectar la precisión y robustez del sistema desarrollado. No se incluye la recolección primaria de datos de audio con hablantes nativos debido a restricciones de tiempo y recursos, aunque esto sería recomendable para investigaciones futuras. Además, el proyecto no abarca el desarrollo de nuevas arquitecturas de modelos de PLN desde cero, sino que se enfoca en la adaptación y optimización de modelos preexistentes para el contexto específico del idioma Bribri.

1.8 Marco de Referencia Legal, Gubernamental y Socioeconómico

De acuerdo con los sitios web actualizados del gobierno se logró obtener una perspectiva más clara del enfoque a nivel organizacional y socioeconómico.

En el caso del Instituto de Desarrollo Rural (INDER), este se comprometió desde el año 2017 a apoyar a los pueblos indígenas Bribri y Cabécares, de forma que va a impulsar iniciativas económicas que respeten las tradiciones indígenas para promover el desarrollo de las comunidades. Además, busca apoyar a personas indígenas que trabajan en el sector agrícola y ganadero, así como promover la infraestructura (puentes, acueductos, remodelación de campos feriales, entre otros esfuerzos).

El Ministerio de educación pública (MEP) posee un departamento llamado Departamento de Educación Intercultural (DEI), el que, según su sitio web, busca promulgar el respeto y fortalecimiento de expresiones culturales diversas, y cuenta con una unidad de educación indígena, donde se tiene como objetivo fortalecer la educación indígena

costarricense, y la elaboración de programas de estudio de lenguas o culturas indígenas. El MEP además expone el marco legal bajo el cual la institución promueve dichas iniciativas, donde existe por ejemplo el *Convenio 169 de la Organización Internacional de Trabajo sobre Pueblos Indígenas y Tribales de 1989*, el que busca el reconocimiento de la cultura e identidad de los pueblos indígenas y tribales, y establece puntos guía para los estados, como el establecimiento de educación bilingüe e intercultural adaptada a la cultura indígena, acceso a la salud, entre otros. En el año 2010 se reformó la ley donde se reforma el Sistema de Educación Indígena de Costa Rica, originalmente establecido en 1993. Esta reforma, hecha por decreto ejecutivo, busca adecuar y actualizar la ley de forma que cumpla con las necesidades actuales de las comunidades indígenas, como, por ejemplo, crea un consejo consultivo nacional de educación indígena, busca aplicar la educación bilingüe, y define una organización administrativo-territorial que se enfoque en las particularidades de cada territorio (anteriormente no se diferenciaban).

El M.E.P. cuenta con varias infografías en su página web dedicada al tema, donde existen recursos educativos para varios pueblos indígenas. En el caso del pueblo Bribri, cuentan con un tomo especial llamado mini enciclopedia donde se detalla la cultura, orden jerárquico, economía, entre otros aspectos de este pueblo. La enciclopedia es bilingüe, y se presenta con códigos de color, en este caso anaranjado para la sección escrita en Bribri y verde para la versión traducida al español.

En los años posteriores a las reformas anteriormente mencionadas, no se han hecho reformas a la ley con respecto a la preservación del lenguaje y la cultura Bribri, exceptuando un decreto ejecutivo en el año 2020 donde el gobierno acató las medidas cautelares impuestas por la corte interamericana de derechos humanos (C.I.D.H.) con respecto a la situación territorial de los pueblos indígenas de Salitre y Telire. En este caso, lo relacionado al lenguaje fue que se crearon infografías para los pueblos indígenas y material audiovisual mostrando protocolos para la prevención del COVID-19. Desde ese entonces no se han tomado acciones a nivel gubernamental para preservar los lenguajes o la cultura indígena.

1.9. Revisión de literatura

1.9.1. Metodología escogida para la revisión sistemática

En cuanto a la metodología para realizar la revisión sistemática de la literatura, se escogió el método Kitchenham, creado por la investigadora Barbara Kitchenham en el año 2007 específicamente para la investigación en el área de ingeniería del software. Conviene subrayar que este método se ha convertido en un estándar para investigaciones desde su introducción, por lo que muchos estudios realizados en el campo utilizan este método para las revisiones sistemáticas.

1.9.2. Planificación de la revisión

1.9.2.1. Pregunta de investigación

¿Cuáles son las técnicas de PLN más efectivas para la transcripción automática de audio a texto en lenguas indígenas de bajos recursos?

1.9.2.2. Palabras clave para la búsqueda

Palabra clave	Traducción al inglés
PLN	NLP
Procesamiento de lenguaje natural	Natural Language Processing
Lenguaje de bajo recurso	Low resource language
Bribri	Bribri
RTA	ASR
Reconocimiento de texto automático	Automatic speech recognition
Transcripción de audio	Audio transcription
Tasa de error por palabra	Word error rate
TEP	WER
Tasa de error por letra	Character error rate
TEL	CER

Tabla 1: palabras clave para revisión sistemática de la literatura

Fuente: elaboración propia

1.9.2.3. Criterios

1.9.2.3.1. Criterios de inclusión

En este caso, se escogerán estudios que sean exclusivamente de los años 2020 en adelante para el tema de procesamiento de lenguaje natural, mientras que para temas fundamentales se pueden referenciar estudios de los años 2010 en adelante.

1.9.2.3.2. Criterios de exclusión

Igualmente, como criterios para excluir estudios para esta revisión, no se tomarán en cuenta estudios que se encuentren fuera del rango de fechas establecido, tampoco estudios que se enfoquen en métodos de NLP para lenguajes de altos recurso. Además, no se tomarán en cuenta estudios que no tengan la implementación práctica o métricas resultantes.

1.9.3. Proceso de búsqueda

Se realizó la búsqueda utilizando la funcionalidad de *advanced search* (búsqueda avanzada) de la plataforma *Google Scholar* (Google académico). Se realizó utilizando las palabras clave y los años de exclusión.

Título del estudio	Ideas centrales	Conclusión
Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri	Separación entre tono y vocales. Prueba sistemática de estilos de transcripción del Bribri	Separar entre el tono y las vocales en las transcripciones puede mejorar la precisión de los modelos en lenguajes de bajos recursos.
Multilingual Models for ASR in Chibchan Languages	Experimentos de ASR con varios modelos (wav2vec2, Whisper, MMS y WavLM)	Wav2vec2 muestra el mejor desempeño de todos los modelos probados. Al probar varios modelos de wav2vec2 para Bribri, el error principal fue debido a la complejidad de los tonos de voz.
An (unhelpful) guide to selecting the right ASR architecture for your under-resourced language	Entrenamiento de varios modelos en lenguajes de bajos recursos de Africa, Asia y América.	El refinamiento de hiper parámetros es clave para mejorar el desempeño de los modelos, ya que todas las arquitecturas probadas mostraron un desempeño similar cuando se probaron con los lenguajes.

Automatic Speech Recognition Advancements for Indigenous Languages of the Americas	Refinamiento de hiper parámetros de modelo wav2vec2 para mejorar tasas de error.	Tasas de error entre 35% y 15% entre 5 lenguajes, incluido el Bribri (35%). La aplicación de el análisis de sensibilidad de Sobol permitió la identificación de hiper parámetros clave para mejorar las tasas de error. Esto los llevó a concluir que cuando hay muy pocos datos, incrementa el riesgo de sobre entrenamiento.
Evaluating self-supervised speech representations for Indigenous American languages	Evaluación de modelos de tipo wav2vec2 entrenados en varios lenguajes, con aprendizaje auto-supervisado (SSL) para el reconocimiento de lenguajes indígenas.	Se probó wav2vec multilenguaje para varios lenguajes, incluyendo el Bribri, pero no se realizó refinamiento de hiper parámetros. Se evaluó el desempeño del modelo en 10 minutos de audio contra 1 hora de audio, con ambos obteniendo altas tasas de error en CER y WER.
Towards an ASR System for Documenting Endangered Languages: A Preliminary Study on Sardinian	Los lenguajes de bajo recurso usualmente no cuentan con muchas transcripciones lo que limita la efectividad de los modelos. Captar características fonéticas de lenguajes	Los audios cortos de aproximadamente 3.5 segundos presentan un reto para los modelos probados, mientras que los audios superiores a 20 segundos mejoran la métrica del CER.

	minoritarios es un reto para los modelos ASR.	El modelo Wav2Vec2 large XLSR-53 mostró buen desempeño. STT Multilingual FastConformer Hybrid también, superando a Wav2Vec2.
Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages	Comparación entre los modelos MMS y XLS-R en 5 lenguajes de bajos recursos. Se necesitan adaptaciones (refinamiento de hiper parámetros) para cada lenguaje debido a la complejidad individual.	MMS muestra mucho potencial cuando la cantidad de datos disponible es minúscula. XLS-R mejora su desempeño cuando existe más de una hora de datos de entrenamiento.
Findings of the Second AmericasNLP Competition on Speech-to-Text Translation	Competencia de tres partes: ASR, traducción de texto y traducción de audio a texto en varios lenguajes incluyendo Bribri.	El enriquecimiento artificial de los datos limitados no genera mejoras en el desempeño de los modelos. Entrenar los modelos con más parámetros mejora el desempeño. El mejor desempeño lo mostró XLS-R wav2vec2. La recolección de audio de lenguajes de bajos recursos es un cuello de botella en la investigación de estos.

Capítulo 2. Marco Conceptual

A continuación, se presentan los conceptos relevantes para esta investigación, que fueron encontrados en la revisión sistemática de la literatura.

2.1. Inteligencia artificial

Según los autores Stuart J. Russel y Peter Norvig, en su libro Artificial Intelligence: A modern approach, “La inteligencia artificial (IA) se ocupa de la creación de sistemas que pueden realizar tareas que normalmente requieren inteligencia humana.” Esta definición abarca desde sistemas simples de reconocimiento de patrones hasta complejas redes neuronales capaces de procesamiento de lenguaje natural y toma de decisiones autónomas.

2.2. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural (PLN) constituye un campo interdisciplinario de la inteligencia artificial que se centra en desarrollar sistemas computacionales capaces de comprender, interpretar y generar lenguaje humano en su forma natural, ya sea hablada o escrita. Según Daniel Jurafsky y James H. Martin, en su libro *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, “el procesamiento de lenguaje natural se define como el campo que se enfoca en proporcionar a las máquinas la capacidad de entender y comunicarse utilizando lenguajes naturales escritos y hablados.”* En este caso, los sistemas de IA basados en PLN integran técnicas computacionales avanzadas, incluyendo algoritmos de aprendizaje automático, modelos estadísticos y arquitecturas de redes neuronales profundas, para analizar la estructura sintáctica y semántica del lenguaje humano, permitiendo así que los programas procesen información textual de manera similar a como lo haría un ser humano.

2.3. Aprendizaje no supervisado

El aprendizaje no supervisado constituye un paradigma fundamental del aprendizaje automático que se caracteriza por la capacidad de los algoritmos de identificar patrones, estructuras y relaciones ocultas en conjuntos de datos que carecen de etiquetas o variables objetivo que se encuentren definidas anteriormente. Según Christopher M. Bishop, en su libro *Pattern Recognition and Machine Learning*, "*el aprendizaje no supervisado es un framework en machine learning donde, en contraste con el aprendizaje supervisado, los algoritmos aprenden patrones exclusivamente a partir de datos no etiquetados*." En este

contexto, los sistemas de inteligencia artificial pueden descubrir automáticamente la estructura de fondo de los datos sin intervención humana directa, utilizando técnicas como *clustering*, reducción de dimensionalidad y detección de anomalías para extraer conocimiento significativo de grandes volúmenes de información no estructurada.

2.4. Preprocesamiento de datos

El preprocesamiento de datos constituye una etapa fundamental en el desarrollo de sistemas de inteligencia artificial y de aprendizaje automático que se encarga de transformar los datos originales a un formato adecuado y utilizable para los algoritmos computacionales. Según Jiawei Han y Micheline Kamber, en su *libro Data Mining: Concepts and Techniques*, *"el preprocesamiento de datos es el proceso de limpieza, integración, transformación y reducción de datos para mejorar la calidad y eficiencia de los algoritmos de minería de datos."* En este proceso, los sistemas implementan técnicas como la limpieza de datos, normalización, manejo de valores faltantes, codificación de variables categóricas y reducción de dimensionalidad, con el objetivo de eliminar inconsistencias, ruido y redundancias que podrían afectar negativamente el rendimiento de los modelos de aprendizaje automático.

2.5. Redes neuronales

Las redes neuronales artificiales constituyen un enfoque computacional inspirado en el cerebro humano, en el que múltiples unidades simples, llamadas neuronas, se conectan entre sí para formar estructuras capaces de aprender patrones complejos a partir de los datos. Según Stuart Russell y Peter Norvig, en su libro *Artificial Intelligence: A Modern Approach*, *“las redes neuronales son redes de unidades de procesamiento simples (neuronas) que pueden aprender funciones complejas ajustando los pesos de las conexiones entre las unidades*” (Russell & Norvig, 2020). En este contexto, las redes neuronales representan una poderosa herramienta para el modelado de relaciones no lineales, con aplicaciones que van desde el reconocimiento de voz hasta la visión por computadora.

2.6. Encoders y decoders

2.6.1. Encoders

El término encoder hace referencia a un componente fundamental en diversas arquitecturas de aprendizaje profundo, especialmente aquellas basadas en modelos

secuenciales o transductores de información. En términos generales, un encoder transforma una entrada compleja —como una secuencia de texto, una señal de audio o una imagen— en una representación latente densa que captura las características semánticas o estructurales más relevantes de dicha entrada. Esta representación intermedia permite que otros módulos del sistema, como los decoders o clasificadores, puedan operar de manera más efectiva sobre los datos transformados.

En el contexto de modelos transformer, según Vaswani et al. (2017), “el encoder es una pila de N capas idénticas, cada una de las cuales tiene dos subcomponentes: una capa de self-attention multi-cabeza y una red feed-forward completamente conectada”. Esta estructura permite que el encoder capture simultáneamente dependencias a largo plazo y representaciones jerárquicas de la entrada, lo cual resulta crucial para tareas como la traducción automática, el análisis de sentimientos o la transcripción de audio.

2.6.2. Decoders

El decoder constituye la segunda mitad en las arquitecturas encoder-decoder y tiene como objetivo generar una secuencia de salida a partir de la representación latente producida por el encoder. En modelos de secuencia a secuencia (seq2seq), el decoder transforma esta representación interna en una secuencia comprensible, como una oración en lenguaje natural, una traducción o una transcripción. Este proceso implica la generación paso a paso de tokens, donde cada paso depende tanto del estado anterior como del contexto global proporcionado por el encoder.

En el caso particular de los Transformers, el decoder está compuesto por múltiples capas que integran mecanismos de atención multi-cabeza tanto sobre la propia secuencia de salida como sobre la representación generada por el encoder. Según Vaswani et al. (2017), “cada capa del decoder contiene tres subcomponentes: una capa de self-attention enmascarada, una capa de atención sobre la salida del encoder, y una red feed-forward completamente conectada”. Esta arquitectura permite que el modelo no solo considere las partes generadas previamente, sino que también incorpore el contexto completo de la entrada, garantizando salidas coherentes y semánticamente relevantes.

2.7. Modelos de transcripción de audio

2.7.1. Wav2Vec2.0

El reconocimiento automático del habla ha sido históricamente dependiente de grandes cantidades de datos etiquetados, lo que ha limitado su escalabilidad y aplicabilidad en contextos con pocos recursos lingüísticos. En respuesta a esta limitación, se han desarrollado modelos de aprendizaje automáticamente supervisado capaces de extraer representaciones significativas directamente del audio sin requerir etiquetas. Una de las contribuciones más influyentes en esta línea es *wav2vec 2.0*, un modelo propuesto por Baevski et al. (2020), que permite aprender representaciones de audio mediante un preentrenamiento contrastivo a partir de grandes volúmenes de datos no etiquetados, seguido de un ajuste fino supervisado con una menor cantidad de ejemplos. Esta arquitectura ha demostrado superar modelos tradicionales tanto en precisión como en eficiencia, marcando un hito en el procesamiento de señales de voz mediante redes neuronales profundas.

2.7.2. Multilingual FastConformer Hybrid

FastConformer constituye una arquitectura mejorada del modelo Conformer original de la empresa Google, diseñada específicamente para lograr un entrenamiento e inferencia eficientes en tareas de procesamiento de voz. Según los investigadores de la empresa NVIDIA que lograron mejorarlo, en su trabajo "Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition", "el modelo propuesto, denominado Fast Conformer (FC), es 2.8 veces más rápido que el Conformer original, soporta escalamiento a parámetros de nivel Billion sin cambios en la arquitectura núcleo y también logra precisión estado del arte en benchmarks de Reconocimiento Automático de Voz". Esta arquitectura implementa un esquema novedoso de submuestreo y reemplaza la atención global con atención de contexto limitado, permitiendo la transcripción de audio de larga duración de hasta 11 horas mientras mantiene alta precisión en diversas tareas de procesamiento de habla.

2.7.3. Whisper

Whisper es un modelo de reconocimiento automático de voz (ASR) de propósito general que representó un avance significativo en el procesamiento del habla mediante técnicas de aprendizaje automático. Según Radford et al., investigadores de la empresa OpenAI, en su investigación *"Robust Speech Recognition via Large-Scale Weak Supervision",* "*Whisper es un sistema de reconocimiento automático de voz entrenado con 680,000 horas de datos supervisados multilingües y multitarea recolectados de la web, demostrando que el uso de un conjunto de datos tan grande y diverso conduce a una mayor robustez ante acentos, ruido de fondo y lenguaje técnico*". Este modelo se fundamenta en una arquitectura de tipo Transformer de secuencia a secuencia que no solo realiza reconocimiento de voz en varios idiomas, sino que también es capaz de ejecutar tareas de traducción de voz e identificación de idiomas, estableciendo un nuevo paradigma en sistemas de procesamiento de audio que pueden generalizar efectivamente a múltiples *benchmarks* estándar sin requerir ajuste fino específico de hiper parámetros para cada conjunto de datos que se procese.

2.7.4. MMS

El modelo MMS (Massively Multilingual Speech) constituye un proyecto de tecnología de habla multilingüe que expande la cobertura de idiomas en sistemas de procesamiento de voz. Según Pratap et al., en su investigación "Scaling Speech Technology to 1000+ Languages", "*el proyecto Massively Multilingual Speech (MMS) aumenta el número de idiomas soportados de 10 a 40 veces, dependiendo de la tarea, siendo los ingredientes principales un nuevo conjunto de datos basado en lecturas de textos religiosos disponibles públicamente y el aprovechamiento efectivo del aprendizaje auto-supervisado*". Este enfoque innovador permite construir modelos wav2vec 2.0 pre entrenados que soportan reconocimiento de voz para más de 1,100 idiomas, modelos de identificación de idiomas capaces de identificar más de 4,000 idiomas (40 veces más que antes), y modelos de texto a voz para más de 1,100 idiomas, democratizando significativamente el acceso a tecnologías de habla para comunidades lingüísticas previamente desatendidas.

2.7.5. WavLM

WavLM se refiere a un modelo de aprendizaje supervisado automáticamente, de gran escala, diseñado específicamente para abordar el procesamiento integral de tareas de habla. Según Chen et al., en su investigación "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing", "*WavLM es un nuevo modelo preentrenado que resuelve tareas de procesamiento de voz de pila completa mediante el aprendizaje conjunto de predicción de habla enmascarada y denoising en el preentrenamiento, lo que permite que WavLM no solo mantenga el modelo de contenido de habla sino que también preserve información multifacética como identidad del hablante, paralingüística y contenido hablado*". Este enfoque innovador permite que el modelo aprenda representaciones universales para todas las tareas de procesamiento de voz, superando las limitaciones de los modelos de aprendizaje auto supervisado anteriores, que se enfocaban principalmente en reconocimiento de voz, y logrando rendimiento estado del arte en el *benchmark* SUPERB con mejoras significativas en diversas tareas de procesamiento de habla.

2.8. Morfología

La morfología constituye una subdisciplina fundamental de la lingüística dedicada al estudio de la estructura interna de las palabras y de los procesos mediante los cuales estas se forman a partir de unidades mínimas dotadas de significado. Según Khanetnok, Srihamongkhon, Daengsaewram y Thabkhoontod (2023), *"la morfología es el estudio de la estructura de las palabras y de las unidades más pequeñas con significado, así como de la manera en que estas se combinan para formar palabras"* (p. 83). En este marco, los lingüistas morfológicos se centran en identificar y analizar morfemas como raíces, prefijos, sufijos y otros elementos y en comprender cómo su combinación da lugar a mecanismos como la derivación, la flexión y la composición. Estos procesos permiten no solo la ampliación del léxico, sino también la adaptación de las palabras a distintas funciones gramaticales dentro de una lengua.

2.9. Fonética

Según Gallinate y Lapierre (2023), en su revista de Lingüística y literatura Página y Signos, *"La fonética, como subdisciplina de la lingüística, se ocupa del estudio de los sonidos del habla para comprender los mecanismos que componen las lenguas en materia del sonido. Los fonetistas pueden dedicarse a la tarea de examinar cómo se producen los sonidos en el tracto vocálico de los hablantes de una lengua. También pueden observar los*

rasgos acústicos y las ondas sonoras que componen un fono (sonido) desde una perspectiva física. Del mismo modo, los especialistas en esta área pueden explicar de qué manera y por qué un hablante de una lengua en específico percibe ciertos sonidos. Estos tres campos de acción corresponden a las principales áreas en las que la fonética se divide tradicionalmente como veremos enseguida."

2.10. Hiper parámetros

Los hiper-parámetros son variables que el investigador fija antes de iniciar el proceso de entrenamiento de un modelo de inteligencia artificial, y que controlan el comportamiento del proceso de aprendizaje, diferenciándose así de los parámetros, que son aprendidos automáticamente a partir de los datos. Según Bischl et al. (2021), en su estudio *Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges*, "la mayoría de algoritmos de aprendizaje automático son configurados por uno o más hiper parámetros que deben ser escogidos muy cuidadosamente y muy frecuentemente tienen un impacto considerable en el desempeño de los modelos". En este sentido, la correcta selección y ajuste de los hiper parámetros es esencial para lograr modelos que generalicen bien, eviten sobreajuste, y alcancen un rendimiento óptimo en tareas como clasificación, regresión o clustering.

2.11. Análisis de Sobol

El análisis de Sobol se define como un método de análisis de sensibilidad, basado en la descomposición de la varianza de la salida del modelo según las variables de entrada o sus combinaciones. Según Fel et al. (2021), en su estudio "*Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis*" los índices de Sobol proporcionan una manera eficiente de capturar interacciones de orden superior a través del lente de la varianza". En este contexto, al aplicar modelos de inteligencia artificial, los índices de Sobol permiten cuantificar la contribución de cada característica, tanto individualmente como en sus interacciones, al comportamiento del modelo de IA. De esta forma, se identifican qué variables tienen mayor impacto sobre la varianza de la salida, lo cual facilita la interpretación de modelos complejos y contribuye a explicar sus decisiones mediante métodos sistemáticos y rigurosos.

Capítulo 3. Marco Metodológico

3.1. Tipo de investigación

Para abordar el desarrollo del sistema de procesamiento de lenguaje natural que permita la transcripción de audio a texto del idioma Bribri, se adopta un enfoque de investigación aplicada con metodología predominantemente cuantitativa.

Esta investigación se enmarca en el paradigma cuantitativo debido a que el proceso de desarrollo, entrenamiento y evaluación del sistema de transcripción se basa en métricas numéricas objetivas y medibles. Como señala Hernández et al. (2014), la investigación cuantitativa utiliza la recolección de datos para probar hipótesis con base en la medición numérica y el análisis estadístico, con el fin establecer pautas de comportamiento y probar teorías.

El carácter aplicado de la investigación busca desarrollar una solución tecnológica específica para la transcripción automática del Bribri. Este enfoque permite la implementación de modelos de inteligencia artificial que pueden ser evaluados mediante métricas cuantificables como el Word Error Rate (WER) y Character Error Rate (CER), proporcionando resultados objetivos y replicables.

Adicionalmente, la investigación incorpora elementos del método experimental, ya que se implementarán y compararán diferentes configuraciones de modelos de NLP, permitiendo establecer relaciones de causa y efecto entre las técnicas aplicadas, los modelos y los resultados obtenidos en la precisión de la transcripción. Este proceso sistemático y controlado facilita la optimización iterativa del sistema propuesto.

3.2. Alcance investigativo

Este proyecto se embarca en un estudio descriptivo y aplicado para desarrollar un sistema de transcripción automática de audio a texto para el idioma Bribri. De acuerdo con Hernández (2014), los estudios descriptivos buscan especificar las propiedades, características y perfiles de personas, grupos, comunidades, procesos, objetos o cualquier otro fenómeno que se someta a un análisis, midiendo o evaluando diversos aspectos del fenómeno a investigar.

Dada la limitada disponibilidad de herramientas tecnológicas para el procesamiento automático de lenguas indígenas costarricenses y la necesidad urgente de preservar el patrimonio lingüístico del pueblo Bribri, esta investigación buscará caracterizar y evaluar el

desempeño de diferentes técnicas de procesamiento de lenguaje natural aplicadas específicamente a esta lengua. El alcance del estudio abarca desde la recopilación y análisis de datos lingüísticos hasta la implementación y evaluación cuantitativa de modelos de transcripción, proporcionando una base tecnológica sólida para futuras investigaciones en el área del procesamiento automático de lenguas indígenas mesoamericanas.

3.3. Enfoque

El enfoque de esta investigación se centrará en un estudio cuantitativo, aplicando metodologías de aprendizaje automático y procesamiento de lenguaje natural para desarrollar y evaluar un sistema de transcripción automática del idioma Bribri. Este enfoque permitirá obtener datos numéricos objetivos y medibles a través de métricas específicas como el Word Error Rate (WER) y Character Error Rate (CER), facilitando un análisis riguroso y sistemático de la efectividad de los diferentes modelos implementados.

La adopción de un enfoque cuantitativo se fundamenta en la naturaleza técnica del problema abordado, donde la precisión y confiabilidad del sistema de transcripción pueden ser evaluadas mediante indicadores estadísticos concretos. Esta metodología permitirá realizar comparaciones objetivas entre diferentes configuraciones de modelos, optimizar parámetros de entrenamiento y validar científicamente los resultados obtenidos, proporcionando evidencia empírica sólida sobre la viabilidad tecnológica de la transcripción automática para lenguas indígenas de baja disponibilidad de recursos.

3.3.1. Justificación de enfoque

El enfoque cuantitativo se justifica por la naturaleza algorítmica del procesamiento de lenguaje natural y la necesidad de evaluar objetivamente el rendimiento de los modelos de transcripción desarrollados. La aplicación exclusiva de métodos cuantitativos permitirá obtener métricas precisas y reproducibles que determinen la efectividad del sistema de transcripción mediante indicadores como WER y CER, así como realizar análisis estadísticos comparativos entre diferentes arquitecturas de modelos y técnicas de preprocesamiento de datos.

Adicionalmente, este enfoque facilitará el establecimiento de métricas cuantificables para la transcripción automática del Bribri que puedan ser replicadas y mejoradas en investigaciones futuras, proporcionando evidencia empírica objetiva sobre la viabilidad técnica de aplicar tecnologías de reconocimiento automático de voz a lenguas indígenas con recursos limitados. La adopción de este enfoque garantiza la rigurosidad científica necesaria para validar los resultados y contribuir al desarrollo de herramientas.

3.4. Diseño

El diseño de esta investigación sigue el enfoque de Investigación Aplicada en Ingeniería de Software, que se centra en el desarrollo y evaluación de sistemas tecnológicos que solucionen problemas específicos mediante la aplicación de técnicas de inteligencia artificial y procesamiento de lenguaje natural. Para lograr esto, se llevarán a cabo varias fases que comprenden desde la recopilación de datos lingüísticos hasta la implementación y evaluación del sistema de transcripción.

La primera fase consistirá en la revisión exhaustiva de literatura especializada utilizando bases de datos académicas como ACL Anthology, IEEE Xplore y arXiv para identificar técnicas actuales de reconocimiento automático de voz aplicadas a lenguas de bajos recursos y establecer el estado del arte en transcripción automática de lenguas indígenas. Posteriormente, se procederá con la recolección y preparación del corpus de datos en idioma Bribri, incluyendo grabaciones de audio y sus correspondientes transcripciones manuales, seguido de procesos de limpieza y normalización de datos para garantizar la calidad del conjunto de entrenamiento.

La fase de implementación abarcará el desarrollo y entrenamiento de múltiples configuraciones de modelos basados en wav2vec 2.0 y otras arquitecturas de NLP, utilizando técnicas de transfer learning y fine-tuning específicas para lenguas de bajos recursos. El análisis de datos se realizará mediante la evaluación cuantitativa de los modelos implementados utilizando métricas estándar como Word Error Rate y Character Error Rate, complementado con análisis estadísticos comparativos para determinar la configuración óptima. Finalmente, la validación de resultados se llevará a cabo mediante pruebas con hablantes nativos del Bribri y la documentación detallada del proceso de desarrollo, contribuyendo así al conocimiento sobre procesamiento automático de lenguas indígenas

mesoamericanas y proporcionando una herramienta práctica para la preservación lingüística.

Capítulo 4. Análisis del diagnóstico

4.1. Comparación de modelos ASR

De acuerdo con la información investigada y recopilada, se tienen 4 candidatos para realizar el proceso de refinamiento de hiper parámetros y de medición respectiva.

Modelo	Ventajas	Desventajas
Wav2Vec 2.0	- Mejores métricas en la mayoría de los estudios consultados - 317 millones de hiper parámetros en su modelo grande - 95 millones de hiper parámetros en su modelo base	- Configuración requiere una muy extensa configuración manual. - Modelo muy grande para GPUs pequeñas. - Poca flexibilidad - Necesita muchos datos de audio y transcripciones.
Whisper	- 39 millones de hiper parámetros en su modelo muy pequeño - 74 millones de parámetros en su modelo base - 244 millones de parámetros en su modelo pequeño	- Amplia cantidad de opciones de modelos. - Modelos pequeños para GPUs pequeñas - Más simple de configurar. - Amplia documentación.

	- 769 millones de parámetros en su modelo mediano - 1.55 miles de millones de parámetros en su modelo grande	- Permite experimentar con varias cantidades de parámetros.
MMS	- Mil millones de parámetros en su modelo tipo encoder del habla - 1.2 miles de millones de parámetros en su modelo decoder de texto	- Modelo muy grande, requiere hardware más potente que un GPU de uso personal - Menor soporte público y documentación
WavLM	- 94 millones de parámetros en su modelo base - 316 millones de parámetros en su modelo grande	- Poca documentación - Requiere ajustes finos importantes para funcionar apropiadamente

4.2. Fuentes de los datos

Figura 1: Corpus pandialectal oral de la lengua bribri

Fuente: Corpus pandialectal oral de la lengua Bribri (2017)

La presente investigación utiliza como fuente primaria de datos el Corpus Pandialectal Oral de la Lengua Bribri (Flores Solórzano, 2017). Este recurso digital constituye una base de datos lingüística fundamental para el desarrollo de sistemas de procesamiento automático del habla en bribri, debido a su cobertura de los tres dialectos principales de esta lengua de tipo chibchense.

El corpus seleccionado se caracteriza por contener registros de habla natural y espontánea, lo cual resulta especialmente valioso para el entrenamiento de modelos de transcripción automática, ya que refleja las variaciones fonéticas presentes en el uso real de la lengua. A diferencia de corpus basados únicamente en textos formales o narrativas tradicionales, esta colección incluye conversaciones cotidianas, monólogos y varios tipos de discurso que proporcionan una muestra representativa del Bribri hablado en contextos naturales.

Para los propósitos de esta investigación, el corpus ofrece ventajas significativas al incluir tanto las grabaciones de audio como sus transcripciones correspondientes, elementos esenciales para el entrenamiento supervisado de modelos de reconocimiento automático de

voz. La disponibilidad de análisis morfológico acompañante enriquece la comprensión de las estructuras lingüísticas del bribri, información que puede ser aprovechada para optimizar el preprocesamiento de datos y mejorar la precisión de los algoritmos de transcripción.

La elección de este corpus responde a criterios de representatividad lingüística, accesibilidad digital y calidad de documentación, factores críticos para garantizar la validez y replicabilidad de los resultados obtenidos en el desarrollo del sistema de transcripción automática propuesto.

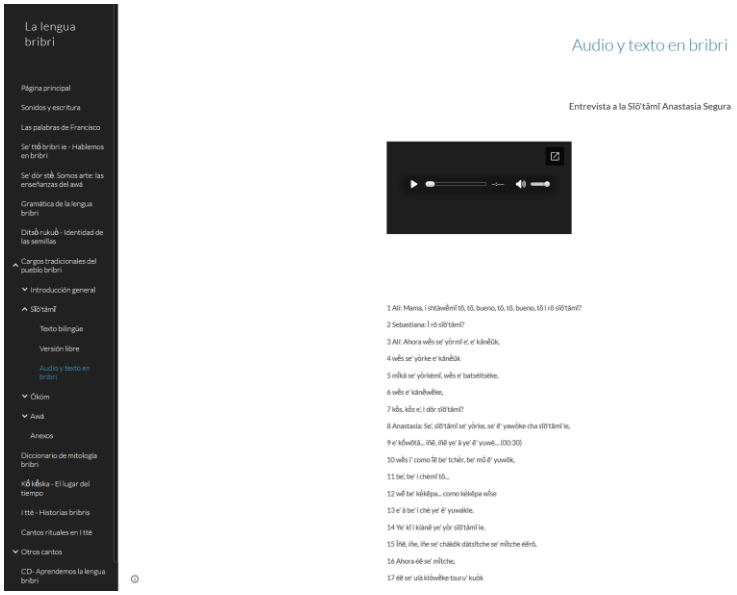

Figura 2: Portal SE'IE La lengua Bribri

Fuente: Portal de la lengua bribri SE'IE, Centro virtual de recursos para el estudio y la promoción de la lengua bribri

Como fuente secundaria de los datos, se utilizó la página SE'IE La Lengua Bribri, iniciativa de la escuela de Filología de la Universidad de Costa Rica. En dicha página se encuentran datos en un formato similar a los del Corpus Bribri Pandialectal Oral, excepto que son mucho más largos en duración ya que son narraciones de historias y conversaciones.

Los materiales audiovisuales disponibles en esta plataforma digital presentan características distintivas que complementan las capacidades del corpus pandialectal. Mientras que el corpus principal contiene segmentos breves de habla espontánea, los archivos de SE'IE documentan narraciones extensas que pueden durar hasta veinte minutos,

permitiendo observar patrones de producción lingüística en contextos de discurso continuo y estructurado. Esta diferencia en la extensión temporal de las grabaciones resulta muy relevante para evaluar el comportamiento del sistema de transcripción automática cuando debe procesar secuencias prolongadas de habla, identificando posibles degradaciones en la precisión conforme aumenta la longitud del audio, ya que permite realizar pruebas con audios largos, así como cortos.

La naturaleza de los contenidos registrados en SE'IE también aporta diversidad tipológica al conjunto de datos de la investigación. Las narraciones incluidas abarcan relatos tradicionales transmitidos por personas mayores de la comunidad Bribri, entrevistas sobre aspectos culturales y ceremoniales, así como conversaciones sobre prácticas cotidianas y conocimientos ancestrales. Esta variedad de géneros discursivos expone al sistema de transcripción a diferentes registros lingüísticos, velocidades de habla y estructuras narrativas, factores que contribuyen a evaluar su robustez y capacidad de generalización ante la heterogeneidad inherente al uso natural de la lengua.

La complementariedad entre ambas fuentes de datos fortalece la fundamentación empírica del proyecto al garantizar que el sistema de transcripción automática sea entrenado y evaluado con muestras representativas tanto de intercambios comunicativos breves como de discursos extensos. Esta estrategia de combinación de recursos lingüísticos responde a la necesidad de desarrollar tecnologías del habla que sean efectivas en múltiples escenarios de uso, maximizando así su potencial de aplicación en contextos educativos, de documentación cultural y de preservación del patrimonio lingüístico inmaterial de la comunidad Bribri.

4.3. Recursos y condiciones disponibles

4.3.1. Infraestructura en la nube

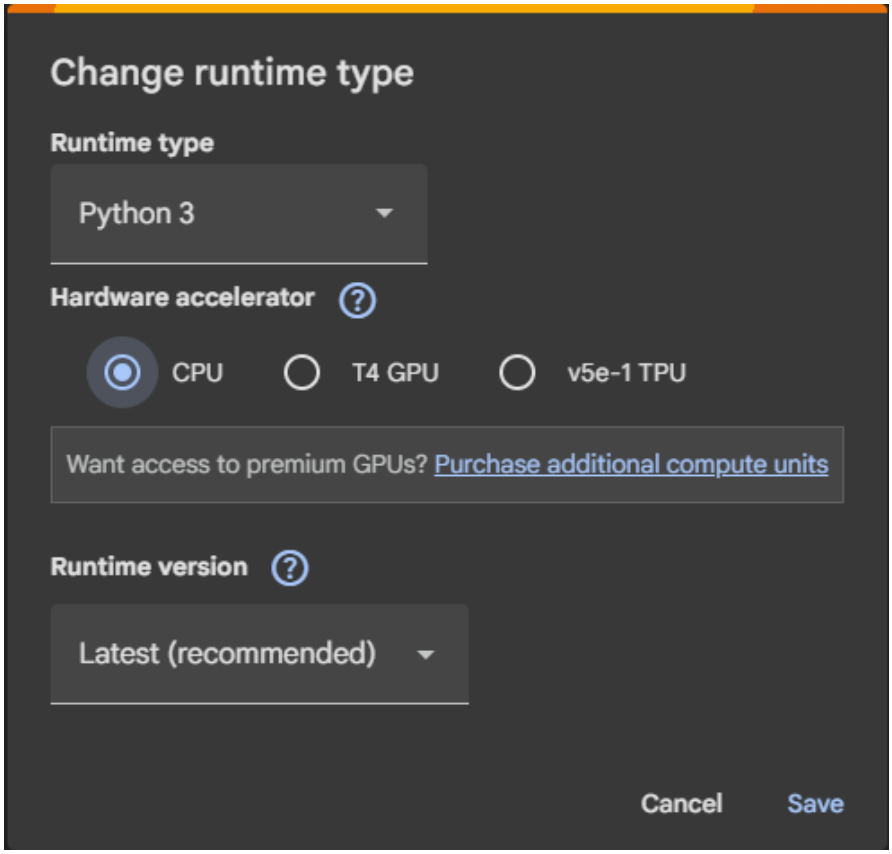

Figura 3: opciones de infraestructura para Google Colaboratory (research.google.com/colaboratory)

Para el desarrollo y experimentación inicial del sistema de transcripción, se utilizará Google Colaboratory en su tier gratuito como plataforma de computación en la nube. Esta plataforma proporciona máquinas virtuales con procesadores Intel Xeon de 2 vCPUs y aproximadamente 13 GB de RAM. En cuanto a aceleración por hardware, el tier gratuito de Colab típicamente asigna GPUs NVIDIA Tesla T4 con 16 GB de VRAM, suficientes para realizar ajustes finos de modelos de tamaño pequeño a mediano.Las sesiones gratuitas de Colab presentan limitaciones importantes que deben considerarse en la planificación del proyecto. Las sesiones pueden ejecutarse por un máximo de 12 horas, dependiendo de la disponibilidad y los patrones de uso, lo que requiere una estrategia de guardado frecuente de checkpoints y gestión eficiente del tiempo de GPU. Esta infraestructura resulta adecuada para la fase de experimentación con modelos más pequeños como Whisper-tiny y Whisper-base, permitiendo validar el pipeline de entrenamiento antes de escalar a recursos más

potentes si fuera necesario. Una ventaja crucial de Colab es su integración nativa con CUDA a través del módulo torch.cuda de PyTorch, que permite aprovechar la capacidad computacional de las GPUs para acelerar los procesos de entrenamiento e inferencia de modelos de aprendizaje profundo. Esta integración facilita la transferencia de tensores y redes neuronales a la GPU sin necesidad de configuraciones complejas, optimizando significativamente los tiempos de experimentación. El entorno viene preconfigurado con los drivers NVIDIA y las versiones compatibles de *CUDA toolki*t, eliminando la necesidad de instalación manual y permitiendo comenzar el desarrollo inmediatamente. La plataforma además ofrece persistencia de datos mediante integración con Google Drive, lo que permite mantener conjuntos de datos, puntos de guardado de modelos y resultados de experimentos entre sesiones. Esto resulta particularmente útil para el desarrollo iterativo del sistema, donde se requiere comparar múltiples configuraciones de hiper parámetros y arquitecturas de modelos. La combinación de recursos gratuitos de GPU con la facilidad de uso y documentación extensa hace de Colab una opción estratégica para la fase inicial de desarrollo, especialmente considerando las limitaciones presupuestarias típicas de proyectos de investigación académica.

4.3.2. Infraestructura local

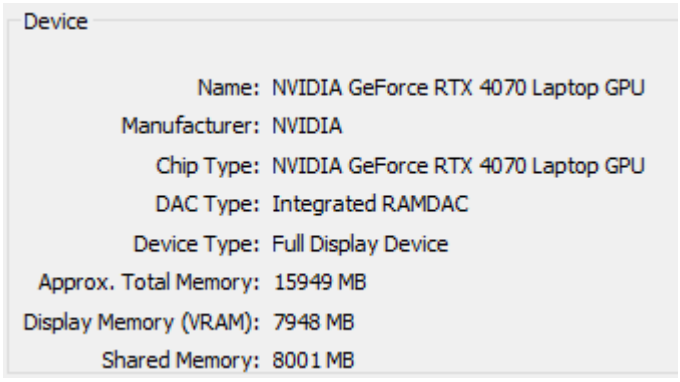

Figura 4: GPU disponible a nivel local

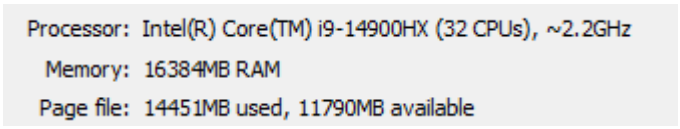

Figura 5: Memoria y procesador disponible a nivel local

Como complemento a la infraestructura en la nube, el proyecto cuenta con un equipo local que servirá para tareas de preprocesamiento de datos, análisis exploratorio, y desarrollo de código. El sistema dispone de un procesador Intel Core i9-14900HX con 24 núcleos y 32 procesadores lógicos operando a 2200 MHz, 16 GB de RAM, un disco de estado sólido de 1 TB para almacenamiento, y una GPU NVIDIA RTX 4070 para laptop con 8 GB de VRAM dedicada.

Esta configuración local permite realizar tareas de preparación de conjuntos de datos, validación de scripts, y experimentación rápida sin depender de la disponibilidad de recursos en la nube. La GPU RTX 4070, aunque con menor memoria que las opciones de Colab, resulta suficiente para inferencia y pruebas rápidas con los modelos más pequeños de Whisper. El procesador i9 de alto rendimiento con sus múltiples núcleos facilita operaciones de preprocesamiento paralelo de grandes volúmenes de datos de audio, mientras que el almacenamiento local agiliza el manejo de los datos de audio y las transcripciones sin necesidad de transferencias constantes a servicios en la nube.

La RTX 4070 cuenta con soporte completo para CUDA, lo que permite aprovechar la aceleración por GPU para todas las operaciones de entrenamiento e inferencia de modelos de aprendizaje profundo de manera local. La tarjeta es

compatible con las versiones más recientes de CUDA toolkit y permite ejecutar PyTorch con aceleración GPU mediante la arquitectura de computación paralela de NVIDIA. Esta capacidad resulta fundamental para el desarrollo iterativo, ya que permite realizar pruebas rápidas de preprocesamiento de audio, validación de transformaciones de datos, y evaluación de modelos sin consumir los límites de tiempo de las sesiones gratuitas de Colab.

El acceso a CUDA en el equipo local complementa estratégicamente la infraestructura en la nube, permitiendo distribuir la carga de trabajo de manera eficiente. Las tareas de preprocesamiento intensivo de audio con librosa, la experimentación con diferentes configuraciones de extracción de características, y las pruebas rápidas de inferencia pueden ejecutarse localmente, mientras que los entrenamientos extensos de modelos más grandes se reservan para las sesiones de Colab con sus GPUs de mayor capacidad. Esta arquitectura híbrida maximiza la productividad del desarrollo, minimiza tiempos de espera, y optimiza el uso de recursos computacionales disponibles. Además, tener un entorno de desarrollo local completamente funcional garantiza continuidad del trabajo incluso cuando los recursos en la nube no están disponibles o han alcanzado sus límites de uso.

4.4. Necesidades y oportunidades

El análisis de los recursos disponibles y las características de los modelos candidatos muestra varias necesidades críticas que deben satisfacerse para el desarrollo exitoso del sistema de transcripción automática. En primer lugar, se requiere un modelo que pueda ejecutarse eficientemente en la infraestructura disponible, específicamente en una GPU NVIDIA RTX 4070 con 8 GB de VRAM. Esta limitación de memoria descarta inmediatamente opciones como el modelo grande de Wav2Vec 2.0 con 317 millones de parámetros, así como los modelos MMS que superan los mil millones de parámetros. La necesidad de flexibilidad en el desarrollo también resulta fundamental, ya que el proceso de experimentación requiere iteraciones rápidas y ajustes constantes de hiper parámetros sin enfrentar configuraciones excesivamente complejas o restrictivas.

Una oportunidad significativa surge de la disponibilidad de la familia de modelos Whisper de la empresa OpenAI, que ofrece una selección completa de

tamaños desde 39 millones hasta 1.55 miles de millones de parámetros. Esta escalabilidad permite comenzar con modelos pequeños para validar el pipeline completo de entrenamiento y evaluación, con la posibilidad de escalar posteriormente a modelos más grandes según los resultados obtenidos y los recursos disponibles. La amplia documentación y el extenso soporte comunitario de Whisper representan otra oportunidad valiosa, ya que reducen significativamente el tiempo de implementación y facilitan la resolución de problemas técnicos durante el desarrollo.

La infraestructura local disponible presenta oportunidades adicionales en términos de control y flexibilidad. El procesador Intel Core i9-14900HX con 24 núcleos permite realizar preprocesamiento paralelo eficiente de grandes volúmenes de datos de audio, mientras que el almacenamiento SSD de 1 TB facilita el manejo ágil del conjunto de datos sin depender de transferencias a servicios externos. La disponibilidad de soporte CUDA completo en la RTX 4070 habilita aceleración por GPU para todas las operaciones críticas del sistema, desde el preprocesamiento de características hasta el entrenamiento e inferencia de modelos.

El proyecto identifica también la oportunidad de establecer un flujo de trabajo completamente local que elimine las limitaciones de las plataformas en la nube gratuitas. Las restricciones de tiempo de sesión de 12 horas en Google Colaboratory, junto con la disponibilidad variable de recursos GPU, representan obstáculos significativos para entrenamientos extensos y experimentación continua. Un ambiente local proporciona sesiones de entrenamiento ininterrumpidas, acceso garantizado a recursos computacionales, y mayor control sobre la configuración del entorno de desarrollo. Esta autonomía resulta particularmente valiosa para proyectos de investigación académica donde la reproducibilidad y el control experimental son prioritarios.

4.5. Conclusión del diagnóstico

El análisis comparativo de modelos ASR y la evaluación de recursos disponibles conducen a conclusiones definitivas sobre la arquitectura y el ambiente de desarrollo del sistema. Después de considerar las ventajas y desventajas de los cuatro candidatos principales, se ha determinado que Whisper representa la opción más viable para este

proyecto. Esta decisión se fundamenta en varios factores técnicos y prácticos que alinean las capacidades del modelo con las limitaciones y necesidades identificadas.

Wav2Vec 2.0, a pesar de demostrar métricas superiores en múltiples estudios donde se aplicó sobre el lenguaje Bribri, presenta obstáculos significativos para su implementación. La configuración extremadamente compleja del modelo requiere una inversión considerable de tiempo en ajustes manuales extensos, mientras que sus 317 millones de parámetros en la versión grande exceden la capacidad práctica de la GPU disponible. La rigidez del modelo y sus altos requerimientos de datos de entrenamiento lo hacen inadecuado para un proyecto con restricciones de recursos computacionales y de datos. De manera similar, MMS y WavLM presentan limitaciones críticas: MMS requiere hardware sustancialmente más potente que el disponible con sus más de mil millones de parámetros, mientras que WavLM sufre de documentación insuficiente y necesita ajustes finos considerables para alcanzar un rendimiento aceptable.

Whisper emerge como la solución óptima al ofrecer una combinación única de flexibilidad, documentación extensa, y opciones de escalabilidad. La disponibilidad de modelos desde 39 millones hasta 1.55 miles de millones de parámetros permite seleccionar variantes que se ajusten perfectamente a la capacidad de la RTX 4070, específicamente los modelos tiny, base, y small. La simplicidad relativa de configuración, combinada con el soporte comunitario robusto y la amplia documentación oficial, reduce significativamente la curva de aprendizaje y acelera el desarrollo. Además, la arquitectura de Whisper permite experimentación eficiente con diferentes tamaños de modelo sin requerir cambios fundamentales en el código o la infraestructura.

En cuanto al ambiente de desarrollo, se ha determinado que el sistema se implementará completamente en la infraestructura local en lugar de depender de Google Colaboratory. Esta decisión responde a la necesidad de mayor control sobre el proceso de entrenamiento y experimentación. Las limitaciones de tiempo de 12 horas por sesión en Colab introducen interrupciones artificiales en entrenamientos que podrían requerir períodos más extensos, forzando estrategias complejas de guardado de checkpoints y reinicio de sesiones. La disponibilidad variable de GPUs en el tier gratuito añade incertidumbre al cronograma de desarrollo, mientras que el ambiente local garantiza acceso consistente y predecible a recursos computacionales.

El equipo local con procesador i9-14900HX, 16 GB de RAM, y GPU RTX 4070 proporciona capacidad suficiente para entrenar y evaluar los modelos pequeños y medianos de Whisper. El control completo sobre el ambiente de desarrollo permite configuraciones personalizadas de software, gestión flexible de dependencias, y debugging profundo sin las restricciones de plataformas en la nube. La persistencia inherente de datos locales elimina la necesidad de sincronización constante con servicios externos, agilizando el flujo de trabajo y reduciendo puntos potenciales de fallo. Adicionalmente, la capacidad de ejecutar experimentos continuos sin límites de tiempo facilita entrenamientos más extensos y búsquedas exhaustivas de hiperparámetros cuando sea necesario.

Esta estrategia de desarrollo local con Whisper como modelo base establece una fundación sólida para el sistema de transcripción automática. La combinación de un modelo bien documentado y flexible con una infraestructura completamente controlada maximiza las probabilidades de éxito del proyecto, permitiendo iteraciones rápidas, experimentación eficiente, y resultados reproducibles. El proyecto procederá con la implementación del pipeline de entrenamiento utilizando las variantes más pequeñas de Whisper para validar la metodología completa, con la posibilidad de escalar a modelos más grandes según lo permitan los resultados iniciales y la capacidad computacional disponible.

Capítulo 5. Desarrollo e implementación

5.1. Arquitectura de la propuesta de solución

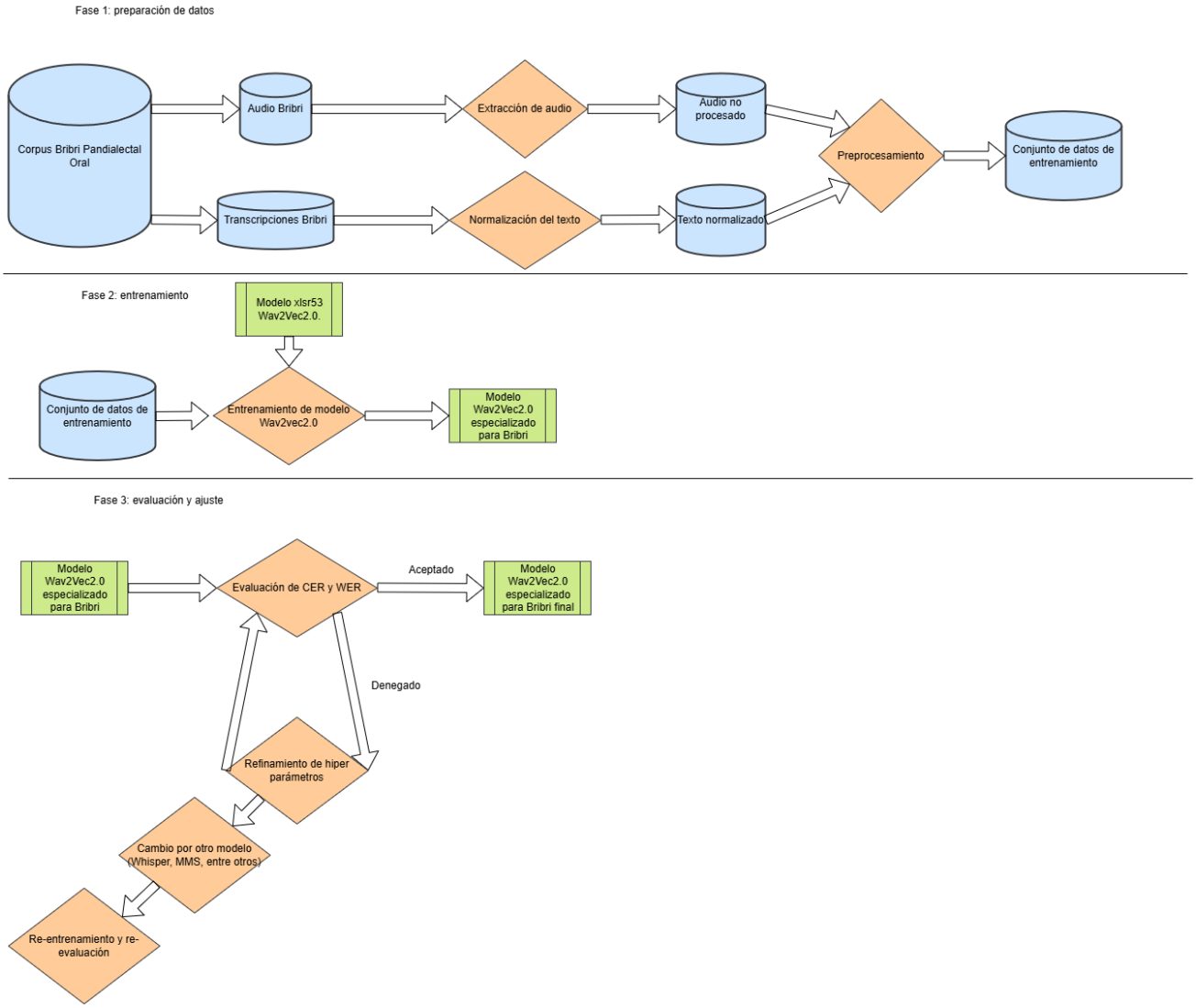

Figura 6: Arquitectura propuesta

Fuente: elaboración propia

La arquitectura de la solución propuesta se organiza en tres fases principales: preparación de datos, entrenamiento del modelo y evaluación y ajuste. A continuación, se detalla cada fase según el flujo de trabajo representado en la Figura 6.

5.1.1. Fase 1: Preparación de datos

En esta primera etapa se parte del Corpus Bribri Pandialectal Oral, el cual contiene tanto archivos de audio en lengua Bribri como sus respectivas transcripciones textuales. A partir de este corpus, se inicia un proceso de preparación de datos que sigue dos líneas paralelas de trabajo. Por un lado, se extrae el audio bruto de los archivos originales y se clasifica como audio no procesado, preservando su calidad original para garantizar la fidelidad de las grabaciones. Por otro lado, las transcripciones en Bribri son sometidas a un proceso de normalización lingüística que permite estandarizar su representación textual, garantizando consistencia ortográfica, coherencia en el uso de caracteres especiales y uniformidad en la segmentación de unidades lingüísticas para su uso posterior en el entrenamiento del modelo.

Además, se incorporan también datos provenientes del sitio web SE'IE La Lengua Bribri, que constituye el resultado del proyecto número 745-B7-7A4 "Centro virtual de recursos para la investigación de la lengua Bribri", desarrollado por la Escuela de Filología, Lingüística y Literatura (EFLL) y el Instituto de Investigaciones Lingüísticas (INIL) de la Universidad de Costa Rica. Esta fuente complementaria aporta grabaciones de mayor extensión temporal y diversidad temática, enriqueciendo el conjunto de datos disponibles para el entrenamiento y validación del sistema.

Una vez que se cuenta con el audio no procesado y el texto normalizado proveniente de ambas fuentes, ambos componentes se integran en una fase de preprocesamiento técnico. En esta etapa se realiza el alineamiento de los pares audio-texto, verificando la correspondencia entre cada segmento de audio y su transcripción asociada, y se formatean según los requerimientos específicos del modelo de entrenamiento seleccionado. Este proceso incluye la conversión de formatos de archivo, la estandarización de tasas de muestreo y la estructuración de los datos en el esquema esperado por la arquitectura del modelo. El resultado final de esta fase es un conjunto de datos de entrenamiento debidamente preparado, validado y organizado para ser utilizado de manera eficiente en la siguiente fase del flujo de trabajo.

5.1.1.1. Recopilación de datos iniciales

Los datos se recopilaron de las fuentes anteriormente mencionadas, las cuales son el Corpus Pandialectal Oral de la Lengua Bribri y el portal SE'IE La Lengua Bribri. Este proceso de recopilación implicó la descarga sistemática de archivos de audio y sus transcripciones correspondientes desde ambas plataformas digitales. Dado que cada fuente presenta características estructurales y organizativas distintas, fue necesario implementar procedimientos específicos de extracción para cada repositorio, documentando meticulosamente el origen de cada archivo para mantener la trazabilidad de los datos a lo largo del proceso de investigación.

5.1.1.1.1. Datos textuales

Figura 7: La lengua Bribri

Página web con datos Bribri

En ambas plataformas, se encontraba inicialmente el material de audio y posteriormente el material de texto asociado. Cabe recalcar que el texto no incluye etiquetas de tiempo, es decir, cada audio se asocia globalmente a su transcripción textual completa, pero no está mapeado el texto al momento específico en el que se pronuncia dentro del audio. Esto significa que se sabe que el texto está asociado al audio en su totalidad, pero no se encuentra segmentado en fragmentos correspondientes a palabras individuales u oraciones específicas con marcas temporales de inicio y fin. Esta característica de los datos

originales representa un desafío metodológico significativo, ya que la ausencia de alineamiento temporal fino entre audio y texto requiere un proceso manual de sincronización que consume tiempo considerable y demanda un conocimiento profundo de la lengua Bribri para identificar correctamente los límites de los segmentos discursivos.

5.1.1.1.2. Datos de audio

Figura 8: Corpus Bribri Pandialectal Oral

Fuente de audio y transcripciones

El audio se encontraba en formatos distintos al requerido por Whisper, entre ellos MP3 y MP4, los cuales debieron ser convertidos al formato WAV estándar esperado por el modelo. Dichos segmentos de audio se encuentran además en tamaños no estandarizados, presentando una variabilidad considerable en su duración. Por ejemplo, algunos archivos tienen una duración superior a los 20 minutos, mientras que otros duran apenas 15 segundos, con una amplia distribución de duraciones intermedias. Esto representa un corpus no estandarizado que requiere un trabajo adicional significativo, ya que los segmentos deben ser procesados y ajustados de forma manual debido a su naturaleza heterogénea. La falta de uniformidad en las duraciones impide la aplicación directa de procesos automatizados de

segmentación, haciendo necesaria la intervención humana para garantizar que cada segmento cumpla con las especificaciones técnicas del modelo de reconocimiento automático de voz.

5.1.1.2. Segmentación de datos manual

Considerando que uno de los requerimientos fundamentales de los datos de entrada para el modelo Whisper es que los segmentos de audio sean de exactamente 30 segundos o menos cada uno, se tomaron todas las muestras recopiladas y se segmentaron manualmente. Este proceso implicó identificar y fragmentar los segmentos de audio superiores a 30 segundos, dividiéndolos en fragmentos de la duración requerida mientras se procuraba respetar los límites naturales del discurso, evitando cortes abruptos en medio de palabras o frases. Este trabajo de segmentación resultó en un total de 279 muestras de 30 segundos cada una, cumpliendo así con el estándar requerido por la arquitectura del modelo Whisper y garantizando la compatibilidad técnica necesaria para el proceso de entrenamiento.

Posteriormente a la segmentación de los fragmentos de audio, se escuchó de forma manual cada segmento de 30 segundos, realizando una labor meticulosa de sincronización entre audio y texto. Las transcripciones, que antes estaban asociadas a audios de más de 20 minutos de duración sin ninguna marca temporal, fueron alineadas manualmente con cada uno de los nuevos segmentos. Se escuchó atentamente cada audio y se delimitó su inicio y su fin en el texto transcrito, asignando a cada fragmento de audio de 30 segundos su correspondiente porción textual específica. Este proceso representó un desafío importante a nivel de escucha manual debido a la complejidad fonológica del idioma Bribri, particularmente en lo que respecta a su sistema tonal y a la presencia de fonemas que no existen en español, requiriendo múltiples reproducciones de cada segmento para garantizar la precisión del alineamiento.

5.1.1.3. Normalización regular del texto

Durante el análisis inicial de las transcripciones se identificó que estas no seguían una convención estándar sobre ciertos caracteres diacríticos y de puntuación. Por ejemplo, los símbolos ' (apóstrofo) y ` (acento grave) se utilizaban de forma intercambiable entre ambos corpus para representar la misma función lingüística, generando inconsistencias en la representación textual. Es por esto que se decidió estandarizar este símbolo y utilizar exclusivamente la comilla simple ' en todas las transcripciones. Esta decisión no implicó

eliminar las tildes vocálicas que forman parte del sistema ortográfico del Bribri, sino únicamente unificar el uso de este símbolo específico que aparece frecuentemente para marcar características tonales o separar morfemas dentro de las palabras. Existen numerosas ocasiones donde este símbolo se usa para delimitar elementos morfológicos o indicar cambios tonales, como por ejemplo en palabras como "se'ie" (como nosotros) o "kő'pa" (casa grande), donde la comilla tiene una función lingüística distintiva.

Al normalizar este símbolo de manera consistente en todo el corpus, se le permitió al modelo aprender patrones lingüísticos con mayor eficacia, partiendo de un corpus estandarizado conjunto que elimina variaciones arbitrarias de notación que podrían confundir el proceso de aprendizaje. Otra normalización realizada consistió en convertir todo el texto a letras minúsculas y asegurar que las palabras estuvieran separadas por un espacio único, ya que el modelo Whisper espera este formato específico de entrada para las transcripciones. Esta conversión a minúsculas simplifica el espacio de vocabulario del modelo y reduce la complejidad del proceso de aprendizaje, dado que elimina la necesidad de distinguir entre versiones mayúsculas y minúsculas de cada carácter.

5.1.1.4 Normalización NFC del texto

Una vez que el texto se encontraba en formato de letras minúsculas, con palabras separadas por un espacio único y con las comillas estandarizadas, se aplicó una normalización adicional a nivel de codificación de caracteres, específicamente relacionada con la Forma de Normalización Canónica Compuesta (NFC, por sus siglas en inglés). En el estándar Unicode, ciertos caracteres con diacríticos pueden ser representados de dos maneras distintas: como un único carácter pre-compuesto o como una secuencia de un carácter base seguido de uno o más caracteres combinantes. Por ejemplo, la vocal "á" puede ser representada internamente como el carácter único U+00E1 o como la secuencia U+0061 *(LATIN SMALL LETTER A)* seguida de U+0301 *(COMBINING ACUTE ACCENT*). Aunque visualmente ambas representaciones se ven idénticas, a nivel de procesamiento computacional son diferentes secuencias de bytes.

La normalización NFC resuelve esta ambigüedad convirtiendo todas las secuencias descompuestas en sus equivalentes pre compuestos cuando estos existen en el estándar Unicode. Este proceso es particularmente importante en el contexto del Bribri, que utiliza diversos diacríticos para representar tonos y características fonológicas específicas. Al aplicar la normalización NFC, se garantiza que cada carácter con diacrítico tenga una

representación única y consistente en todo el corpus, eliminando posibles inconsistencias que podrían surgir si diferentes fuentes o editores utilizaron formas de codificación distintas. Esta estandarización a nivel de codificación permite que el modelo de reconocimiento de voz trate cada carácter de manera consistente durante el entrenamiento y la inferencia.

5.1.1.5. Normalización del audio

Con respecto al procesamiento del audio, se desarrolló un programa auxiliar específico que tomara los segmentos de audio recopilados y los convirtiera a un formato que cumpliera con el estándar técnico esperado por el modelo Whisper. Estos requerimientos técnicos consisten en que el audio debe ser de tipo monoaural (mono) en lugar de estéreo, eliminando así cualquier información de canal dual que no aporta valor para la tarea de reconocimiento de voz y reduce el tamaño de los archivos. Adicionalmente, la frecuencia de muestreo debe ser de 16 kHz, que es suficiente para capturar las frecuencias relevantes del habla humana, y el tamaño de muestra (sample size) debe ser de 16 bits, proporcionando una resolución adecuada para representar la señal de audio sin desperdicio de recursos computacionales.

El programa de normalización implementado procesó automáticamente todos los archivos de audio del corpus, aplicando conversiones de formato cuando fue necesario y verificando que cada archivo cumpliera con las especificaciones requeridas. Este proceso incluyó la conversión desde formatos como MP3 y MP4 al formato WAV estándar, el re-muestreo de frecuencias cuando los archivos originales tenían tasas de muestreo diferentes a 16 kHz, y la conversión de archivos estéreo a monoaural mediante la combinación de ambos canales. La automatización de este proceso garantizó la uniformidad técnica de todos los datos de audio utilizados para el entrenamiento del modelo.

5.1.1.6. Revisión de la calidad del audio

Finalmente, se realizó una revisión manual exhaustiva de la calidad de los audios procesados, con el objetivo de identificar y descartar material de baja calidad acústica o con niveles excesivos de ruido de fondo que pudieran interferir con el proceso de entrenamiento del modelo. Durante esta revisión se escuchó cada segmento evaluando factores como la claridad de la voz del hablante, la presencia de ruidos ambientales, distorsiones o cualquier artefacto que pudiera degradar la señal de habla. También se verificó que no existieran silencios prolongados que ocuparan una porción significativa del segmento de 30 segundos asignado.

Afortunadamente, la calidad general del corpus resultó ser satisfactoria, y este proceso de revisión no resultó en ningún descarte de datos, lo cual indica que las grabaciones originales fueron realizadas con estándares de calidad adecuados. No obstante, la realización de esta verificación manual fue importante para documentar la calidad del corpus y garantizar que todos los datos utilizados para el entrenamiento cumplieran con estándares mínimos de calidad acústica. Este paso de control de calidad contribuye a la validez de los resultados obtenidos posteriormente durante el entrenamiento y evaluación del modelo.

5.1.2. Fase 2: Entrenamiento

Con el conjunto de datos debidamente preprocesado y validado, se procede a la fase de entrenamiento del modelo de reconocimiento automático de voz. En esta etapa se utiliza como punto de partida el modelo Whisper Tiny, desarrollado por OpenAI, el cual representa la versión más compacta de la familia de modelos Whisper. Este modelo ha sido previamente entrenado en un corpus masivo y multilingüe de aproximadamente 680,000 horas de audio etiquetado, recopilado de diversas fuentes de internet. A partir de este modelo pre'entrenado, que ya contiene representaciones generales de características acústicas del habla humana en múltiples idiomas, se realiza un proceso de ajuste fino de hiper parámetros especializado utilizando los datos del corpus Bribri preparado, generando así un modelo Whisper adaptado específicamente para la transcripción automática de la lengua Bribri. Esta estrategia de aprendizaje por transferencia permite aprovechar el conocimiento previo del modelo base y adaptarlo eficientemente a la lengua objetivo, incluso con cantidades limitadas de datos etiquetados.

5.1.2.1. Pruebas iniciales

Se realizaron pruebas iniciales de entrenamiento configuradas manualmente con el objetivo de validar el flujo completo del proceso y verificar la correcta integración de todos los componentes del sistema con la arquitectura Whisper Tiny. Durante estas pruebas exploratorias se identificaron diversos errores de configuración relacionados con la compatibilidad de formatos de datos, la especificación de hiper parámetros específicos del modelo Whisper y la correcta carga de los archivos de audio y texto preprocesados. Estos errores fueron documentados sistemáticamente y se fueron ajustando de forma iterativa mediante modificaciones en los scripts de entrenamiento y en los archivos de configuración.

Este proceso inicial de prueba y error resultó fundamental para comprender el comportamiento del modelo Whisper Tiny con datos en Bribri y para establecer una línea base de rendimiento. Las pruebas manuales permitieron también familiarizarse con los mensajes de error específicos del framework de Hugging Face Transformers utilizado para el entrenamiento, y desarrollar protocolos de solución para problemas comunes que surgieron durante la fase de desarrollo, tales como errores de memoria, incompatibilidades de versiones de bibliotecas y configuraciones incorrectas de los parámetros del tokenizador. Aunque este enfoque manual fue necesario inicialmente, pronto se hizo evidente la necesidad de automatizar el proceso de búsqueda de hiperparámetros óptimos para explorar de manera más eficiente el espacio de configuraciones posibles.

5.1.2.2. Pruebas automatizadas

Para realizar el entrenamiento de forma repetible y sistemática, se implementó el uso de la biblioteca Optuna, un framework de optimización de hiperparámetros que permite ejecutar entrenamientos secuenciales de modelos de inteligencia artificial de manera automatizada. Optuna utiliza algoritmos de optimización bayesiana para explorar inteligentemente el espacio de hiperparámetros, seleccionando configuraciones prometedoras basándose en los resultados de pruebas anteriores. Esta herramienta se utilizó para una fase de automatización temprana que permitió realizar barridos sistemáticos sobre parámetros críticos para el ajuste fino de Whisper Tiny, incluyendo la tasa de aprendizaje, el tamaño de lote, el número de épocas de entrenamiento, la estrategia de congelamiento de capas y diversos parámetros de regularización como el dropout.

La implementación de Optuna permitió ejecutar múltiples experimentos de entrenamiento de manera secuencial sin intervención manual, registrando automáticamente los resultados de cada configuración probada en una base de datos local. Este enfoque automatizado facilitó la identificación de algunos rangos de hiperparámetros útiles que producían mejores resultados en términos de la métrica de evaluación WER (Word Error Rate) sobre el conjunto de validación. Sin embargo, durante esta fase se identificaron limitaciones en cuanto a la visualización de resultados y el monitoreo en tiempo real del proceso de entrenamiento, lo que motivó la búsqueda de herramientas complementarias más robustas para el seguimiento de experimentos y la comparación visual de diferentes configuraciones.

5.1.2.3. Optimización de pruebas automatizadas.

Posteriormente se realizó una migración hacia la plataforma *Weights & Biases* (WandB), que ofrece funcionalidades similares a Optuna para la gestión de experimentos de aprendizaje automático, pero con capacidades superiores de visualización, registro y monitoreo de métricas. WandB permite generar gráficos interactivos del proceso de entrenamiento de los modelos de forma automática y en tiempo real, incluyendo métricas como la función de pérdida durante el entrenamiento, el WER de validación, el uso de memoria GPU y otras estadísticas relevantes para diagnosticar el progreso del ajuste fino de Whisper Tiny. Además, la plataforma ofrece funcionalidades de monitoreo remoto accesible desde cualquier navegador web, lo que facilita el seguimiento de experimentos de larga duración sin necesidad de estar físicamente frente al servidor de entrenamiento o mantener una sesión SSH activa.

Una de las ventajas más significativas de WandB es su capacidad para generar visualizaciones comparativas que permiten analizar el efecto de diferentes hiperparámetros sobre el WER de evaluación obtenido por cada configuración experimental. Estos gráficos de importancia de parámetros, curvas de aprendizaje paralelas y tablas de comparación de experimentos facilitan considerablemente la interpretación de los resultados y la identificación de las configuraciones óptimas para el ajuste fino del modelo. Paralelamente a la adopción de WandB, se realizaron optimizaciones en el propio proceso de entrenamiento para reducir el tiempo total de experimentación y permitir la exploración de un mayor número de configuraciones de hiperparámetros.

Una modificación importante consistió en eliminar la evaluación frecuente del modelo durante el entrenamiento y el cálculo del puntaje de evaluación después de un número pequeño de pasos. Inicialmente, esta evaluación se configuraba para ejecutarse cada 10 pasos de entrenamiento, lo que requería que el modelo se detuviera, ejecutara inferencias sobre todo el conjunto de validación y computara las métricas de evaluación completas, proceso que ralentizaba el entrenamiento de forma considerable debido al overhead computacional adicional. Al eliminar esta evaluación frecuente y reconfigurarla para ejecutarse únicamente al final de cada época completa o cada cierto número significativo de pasos (por ejemplo, cada 500 pasos), se permitió una aceleración importante en las pruebas de entrenamiento y en la búsqueda de hiperparámetros óptimos. Esta optimización redujo el tiempo total de cada experimento sin comprometer significativamente la capacidad de

evaluar el rendimiento del modelo, ya que las evaluaciones periódicas proporcionan información suficiente para monitorear el progreso del entrenamiento y detectar problemas como sobreajuste o estancamiento del aprendizaje.

5.1.3. Fase 3: Evaluación y ajuste

Una vez entrenado el modelo, se evalúa su desempeño utilizando métricas estándar de transcripción automática: CER (*Character Error Rate*) y WER (*Word Error Rate*). Si los resultados obtenidos cumplen con los umbrales de calidad definidos, el modelo se considera aceptado, y se designa como el modelo Whisper especializado para Bribri finalizado.

En caso contrario, se considera denegado, lo que activa una serie de procesos iterativos para mejorar el desempeño. Entre las estrategias consideradas se encuentran el refinamiento de hiper parámetros y el cambio del modelo base (por ejemplo, a alternativas como Whisper, MMS o las previamente definidas en el marco conceptual), y un nuevo ciclo de reentrenamiento y reevaluación.

5.2. Software y herramientas para desarrollo de la solución

Para el desarrollo del sistema de transcripción automática del idioma Bribri, se ha seleccionado un conjunto de herramientas y tecnologías específicas que garantizan tanto la eficiencia en el desarrollo como la calidad en los resultados obtenidos. La selección de estas herramientas se fundamenta en criterios de compatibilidad, rendimiento, soporte comunitario y adecuación a las necesidades específicas del procesamiento de lenguaje natural y aprendizaje automático.

5.2.1. Lenguaje de Programación

La implementación del sistema se realizará utilizando Python como lenguaje de programación principal. Esta elección se justifica por múltiples factores técnicos y metodológicos que lo posicionan como la opción más adecuada para proyectos de inteligencia artificial y procesamiento de lenguaje natural. Como indican Raschka et al. (2020), Python continúa siendo el lenguaje más preferido para computación científica, ciencia de datos y aprendizaje automático, potenciando tanto el rendimiento como la productividad al permitir el uso de bibliotecas de alto y bajo nivel.

La prevalencia de Python en el ecosistema de aprendizaje automático se debe a su extensa colección de bibliotecas especializadas, incluyendo *NumPy* para computación

numérica, *SciPy* para algoritmos científicos, *scikit-learn* para aprendizaje automático clásico, y entre ellas, bibliotecas de aprendizaje profundo como *PyTorch* y *TensorFlow*. Esta riqueza de herramientas disponibles reduce significativamente el tiempo de desarrollo y permite enfocar los esfuerzos en la optimización específica de los modelos para el idioma Bribri, con el objetivo de no volver a implementar funcionalidades básicas.

Adicionalmente, Python ofrece ventajas importantes para la investigación y experimentación, incluyendo su sintaxis clara y legible que facilita el mantenimiento del código, su naturaleza interpretada que permite iteración rápida durante el desarrollo de modelos, y su amplio soporte para visualización de datos y análisis exploratorio, aspectos cruciales para evaluar el rendimiento de los sistemas de transcripción.

5.2.2. Entorno de Desarrollo

Para el desarrollo del código se utilizará *Visual Studio Code* como entorno de desarrollo integrado (IDE). Esta herramienta proporciona un ambiente robusto y versátil que facilita el trabajo con proyectos de aprendizaje automático mediante características específicas como soporte nativo para Python, extensiones especializadas, integración con sistemas de control de versiones, y herramientas de depuración avanzadas.

Visual Studio Code ofrece extensiones específicas para el desarrollo en Python que incluyen autocompletado inteligente, análisis de código estático, formateado automático y soporte para entornos virtuales. Estas funcionalidades son especialmente valiosas en proyectos de procesamiento de lenguaje natural donde la gestión de dependencias y la organización del código son aspectos muy importantes para poder hacer que los experimentos sean replicables.

5.2.3. Bibliotecas Especializadas

El desarrollo del sistema de transcripción se apoyará en bibliotecas especializadas del ecosistema Python para NLP y machine learning. Entre las principales se incluyen PyTorch para la implementación y entrenamiento de modelos de aprendizaje profundo, Transformers de Hugging Face para el acceso a modelos pre-entrenados como Whisper, librosa para el procesamiento de señales de audio, pandas para la manipulación y análisis de datos estructurados, Optuna para la optimización automatizada de hiperparámetros mediante algoritmos bayesianos, y Weights & Biases (WandB) para el monitoreo, visualización y seguimiento de experimentos de entrenamiento en tiempo real.

Esta selección de herramientas garantiza acceso a implementaciones estado del arte de algoritmos de reconocimiento automático de voz, permite la experimentación eficiente con diferentes configuraciones de modelos a través de búsquedas automatizadas de hiperparámetros, facilita el análisis comparativo de resultados experimentales mediante visualizaciones interactivas, y permite la integración de técnicas necesarias para abordar los desafíos específicos de una lengua de bajos recursos como el Bribri.

Capítulo 6. Conclusiones

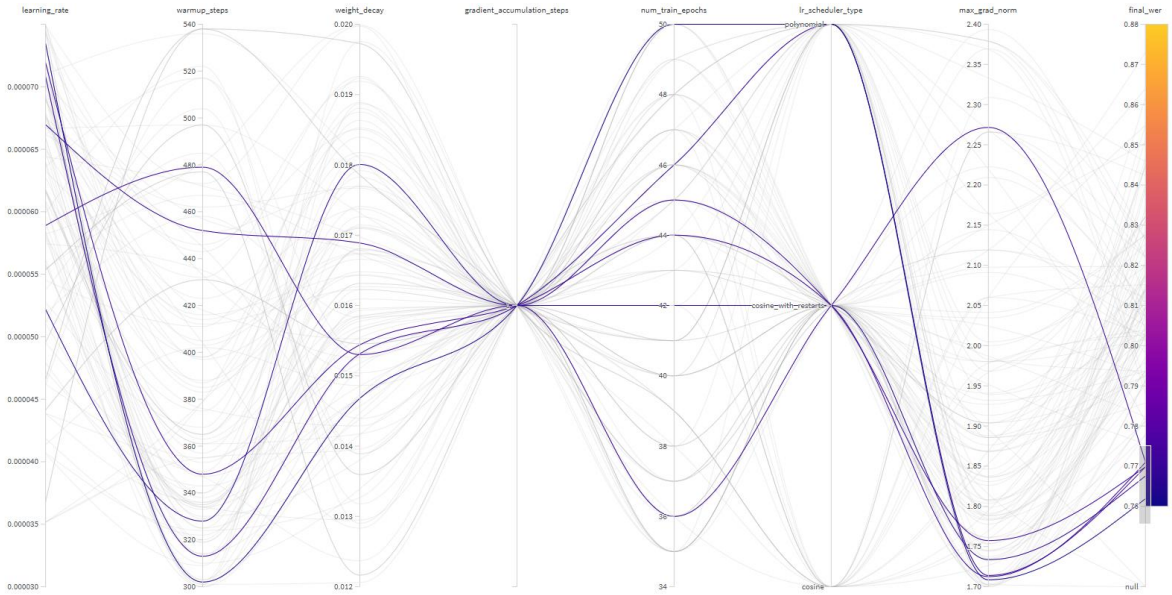

Figura 9: Selección de mejores resultados de las pruebas

Fuente: elaboración propia en plataforma WandB

El presente estudio logró desarrollar un sistema de transcripción automática para la lengua Bribri que supera el estado del arte previamente establecido, demostrando que mediante una cuidadosa preparación de datos y optimización de hiperparámetros es posible alcanzar resultados superiores incluso utilizando modelos más compactos y recursos computacionales significativamente más limitados.

Después de realizar 111 experimentos de entrenamiento utilizando optimización automatizada de hiperparámetros, se identificó la configuración óptima que produjo los mejores resultados. Los hiperparámetros seleccionados fueron los siguientes: una tasa de aprendizaje (learning rate) de 0.00007348, 302 pasos de calentamiento (warmup steps), un factor de decaimiento de pesos (weight decay) de 0.01468, un paso de acumulación de

gradientes (gradient accumulation steps) de 1, un total de 50 épocas de entrenamiento, un programador de tasa de aprendizaje de tipo polinomial (polynomial learning rate scheduler), y una norma máxima de gradiente (maximum gradient norm) de 1.709. Esta configuración específica fue el resultado de un proceso exhaustivo de búsqueda que exploró sistemáticamente el espacio de hiperparámetros mediante las herramientas Optuna y Weights & Biases.

El modelo resultante alcanzó una tasa de error de palabras (WER) de 76.19% en promedio sobre el conjunto de evaluación, lo cual representa una mejora significativa respecto al estado del arte actual establecido por Coto-Solano et al. (2024) en su trabajo "Multilingual Models for ASR in Chibchan Languages" presentado en una conferencia internacional. Este estudio de referencia reportó un WER promedio de 79% utilizando el modelo Whisper Large v2 de OpenAI, que es considerablemente más grande y complejo que el modelo Whisper Tiny empleado en la presente investigación. La mejora de aproximadamente 3 puntos porcentuales en el WER, aunque puede parecer modesta en términos absolutos, representa un avance sustancial considerando las diferencias en los recursos computacionales y el tamaño de los modelos utilizados.

Un aspecto particularmente destacable de los resultados obtenidos es la eficiencia computacional lograda en comparación con el trabajo previo. Coto-Solano et al. (2024) mencionan explícitamente en su estudio las limitaciones relacionadas con los recursos computacionales requeridos, señalando que *"Another important limitation is the amount of computing power needed to train these models. The experiments presented here were exhaustive, but they were also time consuming: The XLSR-53 Wav2Vec2 experiments took 455 GPU hours, using an Nvidia Tesla K80 GPU in a HPC infrastructure. This training was performed in parallel on 5-7 GPUs, and it took approximately one week."* En contraste, el presente estudio logró resultados superiores utilizando únicamente una GPU Nvidia RTX 4070 de 8GB VRAM en una computadora portátil personal, sin acceso a infraestructura de cómputo de alto rendimiento (HPC) ni paralelización en múltiples GPUs. Esta diferencia en los recursos computacionales requeridos demuestra que las optimizaciones implementadas y las técnicas de preparación de datos aplicadas permitieron un uso mucho más eficiente de los recursos disponibles.

En síntesis, se obtuvieron resultados que superan a los reportados por el estado del arte establecido por Coto-Solano et al. (2024) en su trabajo "Multilingual Models for ASR in

Chibchan Languages" presentado en una conferencia internacional de lingüística computacional. Mientras que el estudio de referencia logró un WER promedio de 79% utilizando el modelo Whisper Large v2, el de mayor tamaño de la familia Whisper, y requirió 455 horas de GPU utilizando infraestructura de cómputo de alto rendimiento con múltiples GPUs Nvidia Tesla K80 trabajando en paralelo, el presente estudio alcanzó un WER de 76.19% utilizando el modelo más compacto de la familia, Whisper Tiny, y hardware considerablemente más modesto consistente en una única GPU Nvidia RTX 4070 de 8GB VRAM en una computadora portátil comercial.

Las optimizaciones realizadas durante el proceso de preparación de datos, particularmente la normalización cuidadosa de las transcripciones mediante estandarización de caracteres y normalización NFC, la segmentación manual precisa del audio alineada con el texto correspondiente, y la búsqueda sistemática de hiperparámetros óptimos mediante herramientas de optimización automatizada, fueron factores determinantes que incrementaron significativamente la calidad de los datos de entrenamiento. Estas mejoras metodológicas permitieron que un modelo tan compacto como Whisper Tiny, que contiene aproximadamente 39 millones de parámetros, lograra obtener transcripciones con una tasa de error inferior a la obtenida por modelos sustancialmente más grandes como Whisper Large v2, que contiene más de 1,500 millones de parámetros.

Estos resultados tienen implicaciones importantes para el desarrollo de tecnologías del habla para lenguas de bajos recursos como el Bribri. Demuestran que, con técnicas apropiadas de preparación de datos y optimización de modelos, es posible desarrollar sistemas de reconocimiento automático de voz efectivos sin requerir acceso a infraestructura computacional costosa o modelos de gran escala. Esto reduce significativamente las barreras de entrada para comunidades y grupos de investigación que deseen implementar sistemas similares para otras lenguas indígenas de bajos recursos, democratizando el acceso a estas tecnologías y facilitando los esfuerzos de documentación y revitalización lingüística.

Bibliografía

Fuentes Rodríguez, E. (2011). *Características demográficas y socioeconómicas de las poblaciones indígenas de Costa Rica (Censo 2011)*. Instituto Nacional de Estadística y Censos (INEC), Unidad de Diseño, Procesamiento y Análisis, Área de Censos de Población y Vivienda. https://admin.inec.cr/sites/default/files/media/anpoblaccenso2011-04.pdf_2.pdf

Instituto de Desarrollo Rural. (s.f.). *Breve caracterización del territorio Talamanca-Valle de la Estrella*. https://www.inder.go.cr/talamancavallelaestrella/Caracterizacion-Talamanca-ValleLaEstrella.pdf

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering.* Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University.

Referencias

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., & Boulesteix, A.-L. (2021). Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges. arXiv. https://arxiv.org/abs/2107.05847

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Centro virtual de recursos para el estudio y la promoción de la lengua bribri (SE'IE). (s.f.). Proyecto No. 745-B7-7A4. Escuela de Filología, Lingüística y Literatura e Instituto de Investigaciones Lingüísticas, Universidad de Costa Rica. https://www.lenguabribri.com/

Chen, C.-C., Chen, W., Zevallos, R., & Ortega, J. E. (2023, October 8). Evaluating self-supervised speech representations for Indigenous American languages [Preprint]. arXiv. https://arxiv.org/abs/2310.03639v2

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 2915-2930. https://arxiv.org/pdf/2110.13900

Chizzoni, I., & Vietti, A. (2024). Towards an ASR system for documenting endangered languages: A preliminary study on Sardinian. En CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Pisa, Italy. https://doi.org/10.48550/arXiv

Corpus Bribri. (s.f.). S. Flores Solórzano, F. Morales Morales, y comunidad Bribri, con apoyo de la Universidad de Costa Rica y Cooperación Interuniversitaria UAM-Santander con América Latina. https://bribri.net/index.html

Coto-Solano, R. A. (2021). Explicit tone transcription improves ASR performance in extremely low-resource languages: A case study in Bribri. En Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (pp. 173–184). Association for Computational Linguistics.

Coto-Solano, R., Kim, T. W., Jones, A., & Loáiciga, S. (2024). Multilingual models for ASR in Chibchan languages. En Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 8521–8535). Association for Computational Linguistics.

Deshmukh, P., Kulkarni, N., Kulkarni, S., Manghani, K., & Joshi, R. (2024). Leveraging parameter efficient training methods for low resource text classification: A case study in Marathi. En Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 53-58). https://doi.org/10.1109/I2CT61223.2024.10543946

DIPALICORI: Diversidad y patrimonio lingüístico de Costa Rica. (s.f.). Escuela de Filología, Lingüística y Literatura, Universidad de Costa Rica. https://dipalicori.ucr.ac.cr/

Ebrahimi, A., Mager, M., Wiemerslage, A., Denisov, P., Oncevay, A., Liu, D., Koneru, S., Ugan, E. Y., Li, Z., Niehues, J., Romero, M., Torre, I. G., Alumäe, T., Kong, J., Polezhaev, S., Belousov, Y., Chen, W.-R., Sullivan, P., Adebara, I., ... Kann, K. (2023). Findings of the Second AmericasNLP Competition on Speech-to-Text Translation. In M. Ciccone, G. Stolovitzky, & J. Albrecht (Eds.), Proceedings of Machine Learning Research: Vol. 220. NeurIPS 2022 Competition Track (pp. 217–232). PMLR.

Educación Intercultural. (s.f.). Ministerio de Educación Pública de Costa Rica. https://ddc.mep.go.cr/educacion-intercultural

Fel, T., Cadène, R., Chalvidal, M., Cord, M., Vigouroux, D., & Serre, T. (2021). *Look at the variance! Efficient black-box explanations with Sobol-based sensitivity analysis*. *arXiv*. https://doi.org/10.48550/arXiv.2111.04138

Flores-Solórzano, S. (2019, marzo). La modelización de la morfología verbal bribri.

Flores Solórzano, S., Morales Campos, J., & Morales Morales, F. (s.f.). Analizador Morfológico de la lengua bribri. Universidad de Costa Rica. https://morphology.bribri.net/

Fuentes Rodríguez, E. (2011). Características demográficas y socioeconómicas de las poblaciones indígenas de Costa Rica (Censo 2011). Instituto Nacional de Estadística y Censos (INEC), Unidad de Diseño, Procesamiento y Análisis, Área de Censos de Población y Vivienda. https://admin.inec.cr/sites/default/files/media/anpoblaccenso2011-04.pdf_2.pdf

Galla, C. (2016). Indigenous language revitalization, promotion, and education: Function of digital technology. Computer Assisted Language Learning, 29(7), 1137–1151. https://doi.org/10.1080/09588221.2016.1166137

Gallinate, G. A., & Lapierre, M. (2023). Fonética. En P. Alandia Mercado (Ed.), *Introducción a la lingüística: Curso para investigadores de lenguas indígenas de Bolivia* (1ª ed., pp. 62–99). Página y Signos/Funproeib Andes. https://doi.org/10.5281/zenodo.11106986

Gobierno implementa acciones para la protección de pueblos indígenas. (2020, 9 de junio). Ministerio de Relaciones Exteriores y Culto de Costa Rica. https://www.rree.go.cr/?sec=servicios&cat=prensa&cont=593&id=5568

Han, J., & Kamber, M. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). Metodología de la investigación (6ª ed.). McGraw-Hill.

Instituto de Desarrollo Rural. Breve caracterización del territorio Talamanca-Valle de la Estrella. https://www.inder.go.cr/talamancavallelaestrella/Caracterizacion-Talamanca-ValleLaEstrella.pdf

Jara Murillo, C. V., & García Segura, A. (2018). *Portal de la lengua bribri SE'IE: Centro virtual de recursos para el estudio y la promoción de la lengua bribri* [Sitio web]. Universidad de Costa Rica. https://www.lenguabribri.com

Jimerson, R., Liu, Z., & Prud'hommeaux, E. (2023). An (unhelpful) guide to selecting the right ASR architecture for your under-resourced language. En Proceedings of the 61st

Annual Meeting of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 1008–1016). Association for Computational Linguistics.

Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd ed.). Prentice Hall.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. En Proceedings of the 40th International Conference on Machine Learning (ICML 2023). PMLR. https://arxiv.org/abs/2212.04356

Khanetnok, P., Srihamongkhon, K., Daengsaewram, S., & Thabkhoontod, R. (2023). *Morphology: Word formation in linguistics*. *International Journal of Sociologies and Anthropologies Science Reviews*, 3(1), 83–92. https://doi.org/10.14456/jsasr.2023.9

Liang, S., & Levow, G.-A. (2025). Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2506.17459

Magga, O. H., Nicolaisen, I., Trask, M., Skutnabb-Kangas, T., & Dunbar, R. (2004). Indigenous children's education and indigenous languages. United Nations Permanent Forum on Indigenous Issues.

Ministerio de Educación Pública de Costa Rica. (2017). Minienciclopedia de los bribris y cabécares de Sulá (Tomo 1). https://mep.go.cr/sites/default/files/media/tomo_1.pdf

Organización Internacional del Trabajo. (1989). Convenio sobre pueblos indígenas y tribales, 1989 (Convenio núm. 169). https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@ed_norm/@normes/documents/publication/wcms_100910.pdf

Pratap, V., Wang, C., Adi, Y., Babu, A., Bansal, V., Burchi, M., Elkahky, A., Hsu, W. N., Pino, J. M., Popuri, K., Synnaeve, G., Tomasello, P., Wenzek, G., Williamson, M., Zhang, Y., Zhang, Y., & Dupoux, E. (2023). Scaling speech technology to 1000+ languages. Journal of Machine Learning Research, 25(132), 1-45. https://arxiv.org/abs/2305.13516

Procuraduría General de la República. (2013, 26 de agosto). Reforma del Subsistema de Educación Indígena, Decreto Ejecutivo N.° 37801-MEP. Sistema Costarricense de Información Jurídica.

http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?param1=NRTC&nValor1=1&nValor2=75249&nValor3=93243&strTipM=TC

Rekesh, D., Badger, J., Jain, A., Kuchaiev, B., Huang, T., Nguyen, J., Watanabe, S., & Ginsburg, B. (2023). Fast Conformer with linearly scalable attention for efficient speech recognition. arXiv preprint arXiv:2305.05084. https://arxiv.org/abs/2305.05084

Romero, M., Gómez-Canaval, S., & Torre, I. G. (2024). Automatic speech recognition advancements for Indigenous languages of the Americas. Applied Sciences, 14(15), 6497. https://doi.org/10.3390/app14156497

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193. https://doi.org/10.3390/info11040193

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Solórzano, S. F., & Coto-Solano, R. (2017). Comparison of two forced alignments systems for aligning Bribri speech. Computational Linguistics and Intelligent English Journal, 20(1), 2-15. https://doi.org/10.19153/cleiej.20.1.2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30). https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Zhozhikov, A. V. (2022). Digitalization of the cultural heritage of the indigenous peoples of the Arctic. European Proceedings of Social and Behavioural Sciences, 113, 948–956. https://doi.org/10.15405/epsbs.2022.03.113