

UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

**FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA,
INFORMÁTICA Y MECÁNICA**

**ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE
SISTEMAS**



TESIS

**EVALUACIÓN DE MODELOS DE RECONOCIMIENTO DE VOZ
PARA TRADUCCIÓN QUECHUA-ESPAÑOL EN EL ÁMBITO
MÉDICO EN CUSCO**

PRESENTADO POR:

Br. LUIYI ANTONY QUISPE PALOMINO

**PARA OPTAR AL TÍTULO PROFESIONAL
DE INGENIERO INFORMÁTICO Y DE
SISTEMAS**

ASESOR:

Dr. HARLEY VERA OLIVERA

FINANCIADO POR: ACADEMIA DE
INVESTIGACIÓN Y EDUCACIÓN SUPERIOR
DE BÉLGICA (ARES - UNSAAC)

CUSCO - PERÚ

2026



Universidad Nacional de San Antonio Abad del Cusco

INFORME DE SIMILITUD

(Aprobado por Resolución Nro.CU-321-2025-UNSAAC)

El que suscribe, el Asesor Harley Vera Olivera.....
..... quien aplica el software de detección de similitud al
trabajo de investigación/tesis titulada: Evaluación de Modelos de Reconocimiento
de Voz para Traducción Quechua - Español en el Ámbito Médico en
Cusco.....

Presentado por: Luis Antony Quispe Palomino..... DNI N° 72184526.....;
presentado por: DNI N°:
Para optar el título Profesional/Grado Académico de Ingeniero Informático y
de Sistemas.....

Informo que el trabajo de investigación ha sido sometido a revisión por 02 veces, mediante el
Software de Similitud, conforme al Art. 6° del **Reglamento para Uso del Sistema Detección de Similitud**
en la UNSAAC y de la evaluación de originalidad se tiene un porcentaje de 09 %.

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No sobrepasa el porcentaje aceptado de similitud.	X
Del 11 al 30 %	Devolver al usuario para las subsanaciones.	
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, conforme al reglamento, quien a su vez eleva el informe al Vicerrectorado de Investigación para que tome las acciones correspondientes; Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	

Por tanto, en mi condición de Asesor, firmo el presente informe en señal de conformidad y **adjunto** las primeras páginas del reporte del Sistema de Detección de Similitud.

Cusco, 7 de septiembre..... de 2026.....

f. Harley Vera Olivera
.....
Firma

Post firma.....Harley Vera Olivera.....

Nro. de DNI.....42542815.....

ORCID del Asesor.....0000 - 0003 - 2011 - 8797.....

Se adjunta:

- Reporte generado por el Sistema Antiplagio.
- Enlace del Reporte Generado por el Sistema de Detección de Similitud: oid: 27259:621514775.....

Luiyi Antony Quispe Palomino

Evaluación de modelos de ASR para traducción Quechua Español en el contexto médico cusco.pdf

 Universidad Nacional San Antonio Abad del Cusco

Detalles del documento

Identificador de la entrega

trn:oid:::27259:621514775

Fecha de entrega

7 sep 2026, 9:20 p.m. GMT-3

Fecha de descarga

7 sep 2026, 9:33 p.m. GMT-3

Nombre del archivo

Evaluación de modelos de ASR para traducción Quechua Español en el contexto médico cusco.pdf

Tamaño del archivo

2.5 MB

120 páginas

37.393 palabras

210.177 caracteres

9% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 8 palabras)

Exclusiones

- N.º de coincidencias excluidas

Fuentes principales

- 6%  Fuentes de Internet
- 3%  Publicaciones
- 7%  Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Acrónimos

API	Interfaz de Programación de Aplicaciones (<i>Application Programming Interface</i>)
ASR	Reconocimiento Automático del Habla (<i>Automatic Speech Recognition</i>)
BPE	<i>Byte-Pair Encoding</i>
BLEU	<i>Bilingual Evaluation Understudy</i>
ChrF	<i>Character n-gram F-score</i>
CER	Tasa de Error por Carácter (<i>Character Error Rate</i>)
COMET	<i>Crosslingual Optimized Metric for Evaluation of Translation</i>
CSV	<i>Comma-Separated Values</i>
DL	Aprendizaje Profundo (<i>Deep Learning</i>)
GPU	Unidad de Procesamiento Gráfico (<i>Graphics Processing Unit</i>)
Hz	Hercio (<i>Hertz</i>)
IA	Inteligencia Artificial
INEI	Instituto Nacional de Estadística e Informática
JSON	<i>JavaScript Object Notation</i>
LAAD	Laboratorio de Algoritmos y Análisis de Datos
MFA	<i>Montreal Forced Aligner</i>
ML	Aprendizaje Automático (<i>Machine Learning</i>)
MT	Traducción Automática (<i>Machine Translation</i>)
NLP	<i>Natural Language Processing</i>
NLLB	<i>No Language Left Behind</i>
OOV	Palabras Fuera de Vocabulario (<i>Out-Of-Vocabulary</i>)
PLN	Procesamiento del Lenguaje Natural
S2T	Voz a Texto (<i>Speech To Text</i>)
ST	Traducción de Voz (<i>Speech Translation</i>)
TTS	Texto a Voz (<i>Text To Speech</i>)
WAV	<i>Waveform Audio File Format</i>
WER	Tasa de Error por Palabra (<i>Word Error Rate</i>)

Agradecimientos

Expreso mi más sincero agradecimiento a mi asesor, Dr. Harley Vera Olivera, por su valiosa orientación, acompañamiento constante y confianza brindada durante cada etapa del desarrollo de esta investigación. Sus observaciones y recomendaciones fueron fundamentales para la culminación de este trabajo.

Asimismo, manifiesto mi especial reconocimiento a los profesores Carlos Fernando Montoya Cubas y Gabriela Zúñiga Rojas, integrantes del Laboratorio de Algoritmos y Análisis de Datos (LAAD), por su apoyo académico, disposición para compartir sus conocimientos y valiosos aportes que enriquecieron este proyecto.

De igual manera, agradezco a los 2 colaboradores externos y especialistas en lengua quechua, cuya experiencia, compromiso y colaboración resultaron esenciales para el avance de la presente investigación así como también al técnico en enfermería y transportes Sebastián Edgar Condori Melo por su colaboración en la recolección de datos.

Finalmente, expreso un profundo y especial agradecimiento a mi familia, por su amor incondicional, paciencia, comprensión y constante respaldo en todo momento. Su apoyo ha sido mi mayor motivación para superar cada desafío y continuar con dedicación hasta alcanzar esta meta. Asimismo, agradezco a mis amigos, quienes con su aliento y ánimo permanente contribuyeron a la culminación de este trabajo.

Resumen

En el Perú, el quechua es idioma oficial junto con el español; pese a ello, la población quechuahablante enfrenta una barrera de comunicación en los servicios de salud. Esta investigación evalúa modelos de reconocimiento automático de voz para la traducción de audio en quechua a texto en español en el contexto de consultas médicas. Se recolectaron e integraron distintos conjuntos de datos de audio en quechua con sus traducciones al español, conformando un corpus orientado al entrenamiento y evaluación de modelos de Reconocimiento Automático del Habla (*Automatic Speech Recognition*) (ASR) y Traducción Automática (*Machine Translation*) (MT). Se compararon dos enfoques: sistemas en cascada, formados por un módulo de reconocimiento del habla y otro de traducción automática, y sistemas de traducción directa de voz a texto, evaluados con las métricas *Bilingual Evaluation Understudy* (BLEU) y *Character n-gram F-score* (ChrF). El mejor desempeño correspondió al sistema en cascada con los modelos Whisper y *No Language Left Behind* (NLLB) en su segunda configuración, que alcanzó 5,91 en BLEU y 24,66 en ChrF, superando al sistema de traducción directa basado en SpeechT5. Se concluye que, cuando se dispone de pocos datos de entrenamiento, los enfoques en cascada son una alternativa más efectiva para lenguas de bajos recursos como el quechua que la traducción directa con SpeechT5, la cual requiere mayor cantidad de datos para obtener resultados aceptables.

Palabras clave: Reconocimiento automático del habla, Traducción automática, Consultas médicas, Modelos de lenguaje, Procesamiento del lenguaje natural.

Abstract

In Peru, Quechua is an official language alongside Spanish; nevertheless, the Quechua-speaking population faces a communication barrier in healthcare services. This research evaluates automatic speech recognition models for translating Quechua audio into Spanish text in the context of medical consultations. Different Quechua audio datasets with their corresponding Spanish translations were collected and integrated, forming a corpus for training and evaluating ASR and MT models. Two approaches were compared: cascade systems, composed of a speech recognition module and a machine translation module, and direct speech-to-text translation systems, evaluated using the BLEU and ChrF metrics. The best performance was achieved by the cascade system using the Whisper and NLLB models in their second configuration, which obtained 5.91 in BLEU and 24.66 in ChrF, outperforming the direct translation system based on SpeechT5. It is concluded that, when limited training data are available, cascade approaches are a more effective alternative for low-resource languages such as Quechua than direct translation with SpeechT5, which requires a larger amount of data to achieve acceptable results.

Keywords: Automatic speech recognition, Machine translation, Medical consultations, Language models, Natural language processing.

Índice general

Lista de Tablas	XII
Lista de Figuras	XIII
1. Introducción	1
1.1. El problema	2
1.1.1. Problema General	3
1.1.2. Problemas Específicos	3
1.2. Justificación y motivación	3
1.3. Objetivo General	4
1.3.1. Objetivos Específicos	4
1.4. Alcances y limitaciones	4
1.4.1. Alcances	4
1.4.2. Limitaciones	5
2. Marco Teórico	7
2.1. Quechua	7
2.2. Morfología aglutinante del quechua	10
2.2.1. Tokenización de subpalabras en el quechua	11
2.3. Lenguas de bajos recursos	12
2.4. Aprendizaje Automático	12

2.5. Procesamiento de Lenguaje Natural	13
2.6. Traducción automática	13
2.7. Traducción de voz a texto	14
2.7.1. Enfoque en cascada	14
2.7.2. Enfoque de traducción directa	15
2.7.3. Comparación general	15
2.7.4. Formalización probabilística de los dos enfoques	15
2.7.5. Alucinación en modelos de traducción de voz	17
2.8. Modelo de lenguaje	17
2.8.1. Large Language Models (LLMs)	18
2.8.2. Embedding	18
2.8.3. Tokens	19
2.8.4. Arquitectura Transformer	19
2.8.5. SpeechT5	21
2.8.6. Whisper	23
2.8.7. NLLB	24
2.9. Aprendizaje por transferencia y adaptación de dominio	25
2.10. Métricas de evaluación	25
2.10.1. BLEU	26
2.10.2. ChrF	27
2.10.3. WER y CER	28
2.10.4. Otras métricas: COMET y METEOR	29
2.10.5. Uso complementario de BLEU y ChrF	29
3. Trabajos Relacionados	30
3.1. Revisión de antecedentes	30
3.2. Síntesis y posicionamiento del trabajo	35

4. Metodología	36
4.1. Enfoque y tipo de investigación	36
4.2. Herramientas y entorno de trabajo	36
4.2.1. Google Colab	36
4.2.2. Servicios complementarios de inteligencia artificial	37
4.3. Descripción general de la metodología	37
4.3.1. Diseño de la aplicación	38
4.3.2. Recolección de datos	38
4.3.3. Preprocesamiento	39
4.3.4. Configuraciones experimentales	40
4.3.5. Composición dialectal y variables no controladas	41
4.3.6. Entrenamiento y evaluación	41
4.3.7. Análisis de resultados	42
4.4. Recolección y construcción de datos	43
4.4.1. Fuentes de datos	43
4.4.2. Preprocesamiento de datos	50
4.5. Modelos utilizados	57
4.5.1. ¿Por qué SpeechT5?	57
4.5.2. ¿Por qué Whisper + NLLB?	58
4.5.3. ¿Qué ventaja comparativa tienen?	58
4.5.4. Representaciones vectoriales (embeddings) en los sistemas evaluados	58
4.5.5. ¿Es válida la comparación pese a la diferencia de tamaño entre modelos?	59
4.6. Configuración experimental	61
4.6.1. Diseño experimental	61
4.6.2. Configuraciones evaluadas	61

4.6.3. Experimentos adicionales de aumento de datos y control de fuga	63
4.6.4. Metodología de evaluación	63
4.6.5. Consideraciones experimentales	64
4.7. Declaración sobre el uso de herramientas de inteligencia artificial	64
5. Pruebas y Resultados	66
5.1. Resultados experimentales	66
5.1.1. Consideraciones sobre la normalización de la evaluación	66
5.2. Resultados finales	67
5.3. Análisis de resultados	68
5.3.1. Impacto del dominio médico	69
5.3.2. Atribución del error en el enfoque en cascada	69
5.3.3. Interpretación de BLEU y ChrF	70
5.4. Análisis cualitativo de resultados	71
5.4.1. Matriz de análisis de errores	72
5.5. Discusión	77
5.5.1. Tamaño y calidad de los datos	77
5.5.2. Control de fuga de datos entre entrenamiento y prueba	78
5.5.3. Contrastación con estudios previos	79
5.5.4. Resumen de los resultados obtenidos	79
6. Conclusiones y Trabajos Futuros	81
6.1. Recomendaciones	83
6.2. Trabajos futuros	84
Apéndices	85
A. Imágenes de la Aplicación	86

A.1. Interfaz de login	86
A.2. Pantalla de grabación de audios	87
A.3. Pantalla de validación de audios	88
A.4. Pantalla de registro del entrevistado y consentimientos	89
B. Estructura del Conjunto de Datos	91
B.1. Organización de Carpetas	91
C. Repositorios del Proyecto	93
C.1. Aplicación Android Recolector de Audios en Quechua	93
C.2. Backend Recolector de Audios en Quechua	93
D. Traducciones generadas en el conjunto de prueba del dataset Médico	94
D.1. Traducciones del modelo SpeechT5 entrenado en la configuración 2 . . .	94
D.2. Transcripciones y traducciones del sistema cascada entrenado en la configuración 2	95
E. Alineación y segmentación	98
E.1. Archivo TextGrid de texto alineado	98
F. Experimentos de Ablación	99
F.1. Sistema directo (SpeechT5): aumento de audio y dominio médico . . .	99
F.2. Componente ASR (Whisper): SpecAugment	100
F.3. Componente de traducción (NLLB): escala, curación y oversampling . .	101
F.3.1. Control de dominio general (FLORES-200)	101
Bibliografía	107

Índice de tablas

2.1.	<i>Condición de vitalidad de las variedades del quechua en el Perú</i>	9
2.2.	<i>Ejemplo de sufijación a partir de la raíz wasi (“casa”)</i>	10
2.3.	<i>Fragmentación de palabras quechua bajo el tokenizador de NLLB (SentencePiece)</i>	11
4.1.	<i>Distribución por rango de duración de los audios</i>	52
4.2.	<i>Comparación de métricas BLEU y ChrF entre distintos sistemas de traducción</i>	55
4.3.	<i>Representaciones vectoriales empleadas por cada modelo y su tratamiento durante el ajuste fino</i>	59
4.4.	<i>Número de parámetros de los modelos preentrenados empleados en cada enfoque</i>	60
5.1.	<i>Resultados obtenidos – Configuración 1</i>	67
5.2.	<i>Resultados obtenidos – Configuración 2</i>	67
5.3.	<i>Resultados de evaluación en el dataset laqla – Configuración 1</i>	68
5.4.	<i>Resultados de evaluación en el dataset laqla – Configuración 2</i>	68
5.5.	<i>Atribución del error en el enfoque en cascada – dataset Médico, Configuración 2</i>	69
5.6.	<i>Matriz de análisis de errores en el conjunto de prueba médico (Configuración 2)</i>	73
5.7.	<i>Control de fuga: desempeño del sistema en cascada con y sin deduplicación del corpus – dataset Médico (n = 178), configuración de máximo esfuerzo</i>	78

5.8. <i>Resumen de los resultados obtenidos</i>	80
F.1. <i>Ablación del sistema directo (SpeechT5) — test médico</i>	99
F.2. <i>Ablación de SpecAugment en el componente ASR (Whisper) — validación médica</i>	101
F.3. <i>Ablación del componente de traducción (NLLB) — validación médica</i> .	101
F.4. <i>Evaluación en FLORES-200 (dominio general): base vs. fine-tuned</i> . .	102

Índice de figuras

2.1. Arquitectura transformer	20
2.2. Arquitectura del modelo SpeechT5	22
2.3. Arquitectura del modelo Whisper	23
4.1. Diagrama de flujo de la metodología	38
4.2. Arquitectura del sistema de recolección de datos	48
4.3. Flujo de limpieza de datos de texto	50
4.4. Estructura del corpus de audio	53
4.5. Generación de diccionario fonético	53
4.6. Pipeline de entrenamiento Whisper + NLLB	62
4.7. Pipeline de entrenamiento Speeclt5	62
A.1. Pantalla de login de la aplicación.	87
A.2. Interfaz de grabación de audios.	88
A.3. Interfaz de validación de audios.	89
A.4. Pantalla de registro de datos no identificativos del entrevistado, con los dos puntos de verificación de consentimiento informado (consentimiento por documento firmado en PDF y consentimiento rápido registrado por voz).	90
E.1. Archivo TextGrid de texto alineado.	98

Capítulo 1

Introducción

El quechua, reconocido como lengua oficial del Perú junto con el español según el artículo 48 de la Constitución Política, constituye una parte esencial del patrimonio lingüístico y cultural del país. Según el Censo Nacional 2017 realizado por el Instituto Nacional de Estadística e Informática (INEI), aproximadamente el 13.9 % de la población, equivalente a más de 3.7 millones de personas, tiene al quechua como lengua materna ([Instituto Nacional de Estadística e Informática, 2018](#)). Sin embargo, pese a esta representativa comunidad de hablantes, persisten brechas significativas en el acceso a servicios públicos y privados en su idioma nativo, especialmente en sectores críticos como la salud. La ausencia de atención médica en la lengua materna de los pacientes puede generar barreras comunicativas que dificulten diagnósticos precisos y tratamientos adecuados, debido a que es necesario que un paciente pueda indicar claramente sus síntomas para que un doctor pueda brindar un diagnóstico preciso, afectando directamente la calidad de la atención médica.

Ante este panorama, los avances tecnológicos en el campo de la Inteligencia Artificial (IA) y, en particular, del Procesamiento del Lenguaje Natural (PLN), ofrecen nuevas oportunidades para abordar las limitaciones lingüísticas ya que en la actualidad existen modelos muy capaces en el procesamiento del lenguaje y áreas específicas como el ASR y MT. Herramientas basadas en ASR y MT han mostrado resultados prometedores en la eliminación de barreras idiomáticas en diversos contextos y lenguas. No obstante, el desarrollo de estos modelos para el quechua aún enfrenta desafíos significativos debido a la escasez de datos de calidad, especialmente en dominios especializados como el de la salud ([Moreno Veliz, 2021](#)).

Los modelos más actuales presentan una deficiencia en el tratamiento de idiomas cuya presencia en los datasets de entrenamiento no fue muy significativa, lo que implica que en el entrenamiento no vieron suficientes datos para aprender todas las características necesarias para procesar estos idiomas con buenos resultados, un ejemplo de este tipo de idiomas subrepresentados en los datasets de entrenamiento es el quechua que no tiene mucha presencia en internet a diferencia de otros idiomas como el español, inglés o chino.

La falta de datasets amplios y representativos limita el entrenamiento de modelos de aprendizaje profundo capaces de reconocer y traducir eficazmente el quechua en contextos reales. A esto se suma la diversidad dialectal del idioma, que presenta variaciones fonéticas y léxicas relevantes según la región, debido a que el quechua no es hablado de igual forma en los diferentes países sudamericanos o incluso en los diferentes departamentos del Perú. En ese sentido, el presente trabajo delimita su ámbito de estudio a la variante Collao, en los términos que se detallan en la Sección 1.4.

Considerando la importancia del quechua en el país, la brecha que la comunidad quechuahablante enfrenta en el servicio de salud y el estado actual de los modelos de lenguaje más recientes con respecto al procesamiento de idiomas de bajos recursos, se propone dar los primeros pasos en la implementación de una herramienta de traducción de audio en quechua a texto en español. Para este tipo de traducción de audio a texto existen 2 enfoques, el enfoque de traducción en cascada el cual implica un proceso de 2 pasos (transcripción y traducción) y el enfoque de traducción directa para el que es necesario un modelo multimodal que pueda manejar tanto audios como texto, cada uno de estos enfoques presenta sus ventajas y desventajas que se explicarán más adelante.

1.1. El problema

La población quechuahablante en el Perú enfrenta una brecha en el servicio de salud a causa de un problema de comunicación que puede surgir por la diferencia de los idiomas hablados por parte del doctor como del paciente, conllevando esto a un servicio no equitativo a este sector de la población debido a la importancia de la comunicación entre el paciente y el doctor para brindar un servicio de calidad. Entonces, ¿cómo podemos ayudar a reducir la brecha utilizando la tecnología actual para dar los pasos iniciales en la implementación de una herramienta de traducción de audios en quechua a texto en español? Existen modelos muy capaces en el procesamiento de texto y audio que se pueden combinar en un sistema de 2 etapas así como usar de forma individual para una traducción directa en la tarea de traducción de audio a texto; sin embargo, presentan dificultades frente a idiomas de bajos recursos.

Actualmente, los modelos de ASR y de MT han alcanzado avances significativos en idiomas con grandes volúmenes de datos disponibles, como el inglés y el español. Sin embargo, estos modelos suelen ser entrenados principalmente con lenguas de amplia cobertura y alta disponibilidad de corpus, por lo que muchos de ellos no han sido expuestos a suficientes datos de idiomas de bajos recursos, como el quechua, durante su proceso de entrenamiento ([Moreno Veliz, 2021](#)). Esta situación repercute en tareas de reconocimiento y traducción automática, ya que los modelos poseen una representación limitada de las características lingüísticas de dichas lenguas, además de encontrarse limitados a vocabularios de ciertos idiomas en algunos casos, como ocurre con determinados modelos preentrenados de procesamiento de voz.

Asimismo, la diversidad dialectal del quechua representa un reto adicional, dado que no todas sus variantes poseen el mismo nivel de documentación ni disponibilidad

de recursos lingüísticos. La delimitación dialectal adoptada en este estudio, así como sus implicaciones sobre la interpretación de los resultados, se declaran en la Sección 1.4.

1.1.1. Problema General

¿Cuál es el desempeño de los modelos de reconocimiento de voz y traducción automática, evaluados sobre un corpus del dominio médico, para la traducción de audios en quechua a texto en español en el ámbito médico de la región Cusco?

1.1.2. Problemas Específicos

1. ¿Cómo construir un corpus inicial de pares audio–texto, conformado por grabaciones en quechua producidas por hablantes nativos y sus correspondientes textos en español del dominio médico?
2. ¿Qué técnicas de preprocesamiento son necesarias para limpiar, normalizar y estructurar los datos recolectados e integrarlos con datasets existentes para el entrenamiento de modelos de reconocimiento de voz?
3. ¿Qué configuraciones de datos deben establecerse para comparar de forma controlada y equitativa el enfoque en cascada y el de traducción directa en la traducción de voz quechua–español del dominio médico?

1.2. Justificación y motivación

La comunicación efectiva es un pilar fundamental cuando se trata de brindar servicios de salud de calidad, garantizando que los pacientes puedan expresar sus síntomas y recibir diagnósticos precisos. Sin embargo, en la región del Cusco, donde una porción significativa de la población tiene el quechua como idioma nativo, la barrera lingüística entre los pacientes quechuahablantes y el personal de salud sigue representando un desafío importante. La ausencia de herramientas adecuadas para la traducción del quechua al español en entornos médicos limita el acceso equitativo a la atención sanitaria, lo que puede derivar en diagnósticos incorrectos y tratamientos ineficaces.

A pesar del progreso en los nuevos métodos de procesamiento del lenguaje natural, el ASR en quechua sigue siendo un área con recursos limitados, especialmente en el ámbito médico. Existen pocos datasets especializados en el dominio médico en quechua, lo que dificulta la implementación y la evaluación de modelos de ASR eficientes en este contexto. Otro factor a considerar es que la diversidad dialectal del quechua plantea un reto adicional, ya que no todas sus variantes cuentan con la misma documentación y recursos disponibles.

Este estudio contribuirá al desarrollo de tecnologías para lenguas originarias y fomentará la preservación del quechua en el ámbito digital. En un contexto donde la tecnología es clave para reducir desigualdades, esta investigación representa un paso hacia la inclusión lingüística en la salud, promoviendo soluciones creativas que optimicen la calidad de vida de las comunidades quechuahablantes en el Perú.

1.3. Objetivo General

Evaluar el desempeño de los modelos de reconocimiento de voz y traducción automática sobre un corpus del dominio médico, para determinar cuál presenta mejor desempeño en la traducción de audios en quechua a texto en español en el ámbito médico de la región Cusco.

1.3.1. Objetivos Específicos

1. Recolectar pares de datos de audios en quechua y sus respectivas traducciones al español, utilizando un sistema de recolección de datos, mediante la lectura de textos diseñados en el dominio médico por hablantes nativos de quechua.
2. Aplicar técnicas de preprocesamiento para limpiar, normalizar y estructurar los datos recolectados, integrándolos con datasets existentes para la evaluación de los modelos.
3. Definir las configuraciones experimentales de datos, la agrupación de los corpus en los conjuntos de entrenamiento, validación y prueba, que permitan comparar de forma controlada y equitativa el enfoque en cascada y el de traducción directa.

1.4. Alcances y limitaciones

En esta sección se delimitan los alcances y limitaciones del trabajo de investigación, que deben tenerse en cuenta para la interpretación de los resultados.

1.4.1. Alcances

El trabajo se centra en los siguientes aspectos:

- Traducción automática de voz en un solo sentido, audio en quechua como entrada y texto en español como salida. No se aborda la dirección inversa ni la generación de voz sintetizada en la salida.

- Ámbito médico, en el registro propio de las preguntas y respuestas de una consulta clínica. Los resultados no se extienden a otros dominios especializados.
- El trabajo se enfoca en el Quechua Collao, predominante en la región del Cusco, que es la variante que aportan los corpus qcc, Médico y LAQLA y sobre la cual se evalúan los sistemas. El corpus Siminchik incorpora además la variante Chanka del quechua sureño.
- Dos enfoques de traducción de voz a texto, el enfoque en cascada y el enfoque de traducción directa. No se evalúan otras arquitecturas ni otros modelos preentrenados.
- Evaluación automática y de extremo a extremo, mediante las métricas BLEU y ChrF calculadas sobre el texto en español producido por los sistemas de traducción completos, complementadas con Tasa de Error por Palabra (*Word Error Rate*) (WER) y Tasa de Error por Carácter (*Character Error Rate*) (CER) para el análisis aislado del componente de reconocimiento de voz.

Quedan fuera del alcance de esta investigación el desarrollo de un sistema desplegable en un establecimiento de salud, la validación clínica de las traducciones generadas, el procesamiento en tiempo real y la traducción en la dirección español a quechua.

1.4.2. Limitaciones

La principal limitación del estudio es el tamaño del corpus del dominio médico, de aproximadamente 4 horas de grabaciones. Esta cifra corresponde únicamente al corpus clínico y no al total de datos de entrenamiento, el componente de reconocimiento de voz se ajusta sobre alrededor de 42 horas de audio en quechua, mientras que el componente de traducción parte de un modelo multilingüe preentrenado y se ajusta sobre cientos de miles de pares de texto. El sistema no aprende, por tanto, el quechua a partir de solo 4 horas, estas aportan la especialización terminológica del dominio médico.

Las grabaciones del dataset Médico se obtuvieron mediante la lectura de frases previamente definidas, por lo que el corpus recoge la pronunciación propia de cada participante pero no la expresión espontánea de una consulta real. El corpus Siminchik, procedente de transmisiones radiales, presenta ruido y música de fondo, inconsistencias en la ortografía resultado de la falta de estandarización escrita del quechua. Además, sus traducciones al español se generaron de forma automática y no por expertos en quechua, por lo que constituyen una referencia aproximada y no una versión exenta de errores.

La variante dialectal constituye una variable no controlada del diseño experimental, el entrenamiento incluye una mezcla de Chanka y Collao a través de Siminchik, mientras que la evaluación se realiza exclusivamente sobre corpus de variante Collao. Esta decisión se asume de forma consciente y responde a la escasez de recursos disponibles.

En consecuencia, los resultados no garantizan el mismo nivel de desempeño en otras variantes del quechua ni en contextos monodialectales estrictos.

Los modelos empleados fueron desarrollados mayoritariamente sobre lenguas de altos recursos y su vocabulario no contempla de forma adecuada el quechua. En el caso particular de SpeechT5, el tokenizador no incluye los caracteres acentuados ni la letra “ñ”, lo que obligó a adoptar una normalización ortográfica común para que la comparación entre ambos enfoques resultara metodológicamente equivalente.

Las métricas automáticas empleadas ofrecen una medición cuantitativa y reproducible, pero no capturan por completo los aspectos semánticos del lenguaje en contextos médicos reales, por lo que no sustituyen una validación humana realizada por especialistas bilingües.

Las cifras de BLEU publicadas por otros trabajos sobre traducción de quechua a español no son directamente comparables con las de esta tesis, dado que se calculan sobre conjuntos de prueba, dominios y variantes dialectales distintos. Una comparación rigurosa exigiría evaluar todos los sistemas sobre un mismo conjunto de prueba, lo cual se realiza en el presente trabajo sin embargo no se realiza en los trabajos relacionados.

Capítulo 2

Marco Teórico

Este capítulo establece los fundamentos teóricos sobre los que se apoya la tarea de traducción de voz del quechua al español en el dominio médico. Se inicia con la caracterización del quechua y de su morfología aglutinante, cuyo efecto sobre la tokenización de subpalabras se cuantifica sobre el propio corpus. A continuación se delimita la noción de lengua de bajos recursos y se presentan los fundamentos del aprendizaje automático y del procesamiento de lenguaje natural. Seguidamente se formalizan de manera probabilística los dos enfoques de traducción de voz en cascada y directo y se describen las arquitecturas neuronales empleadas: Transformer, SpeechT5, Whisper y NLLB. Finalmente se definen las métricas de evaluación y los fundamentos del aprendizaje por transferencia y la adaptación de dominio, conceptos sobre los que se sustenta el análisis de resultados del Capítulo 5.

2.1. Quechua

El quechua es uno de los idiomas autóctonos más hablados de América del Sur y tiene una gran importancia histórica y cultural. Su distribución actual en el Perú es el resultado de un proceso histórico de expansión y diversificación en diversas variedades geográficas. A pesar de su relevancia, muchas de sus variantes se encuentran en peligro de desaparición. Este marco teórico aborda la clasificación, distribución, vitalidad y escritura del quechua, así como su rol en la sociedad actual y los retos que enfrenta.

El quechua pertenece a una familia lingüística que se extiende por siete países de Sudamérica: Perú, Ecuador, Colombia, Argentina, Bolivia, Chile y Brasil ([Ministerio de Educación, 2013](#)). En el Perú, se distinguen dos grandes ramas:

- **Quechua I:** Se ubica en la zona central del país y es considerado el más antiguo.
- **Quechua II:** Se divide en las variedades del sur y norte, resultado de procesos históricos de expansión y desplazamiento lingüístico.

Dentro de estas ramas, el Ministerio de Educación ([Ministerio de Educación, 2013](#)) establece cuatro grandes grupos:

1. **Quechua Amazónico:** Incluye el Kichwa amazónico, hablado en Madre de Dios, Loreto, Chachapoyas y San Martín.
2. **Quechua Norteño:** Agrupa variedades como el Quechua Cajamarca (seriamente en peligro) y el Quechua Inkawasi Kañaris (vital).
3. **Quechua Central:** Comprende el Quechua Pataz (vital) y otras variedades en riesgo, como el Quechua Yauyas y el Quechua Wanka
4. **Quechua Sureño:** Incluye variedades como el Quechua Chanka y el Quechua Collao, ambos considerados vitales en varias regiones.

De acuerdo con el Ministerio de Educación ([Ministerio de Educación, 2013](#)), el quechua es un idioma en situación de vitalidad. Sin embargo, ha experimentado una significativa pérdida de hablantes. A inicios del siglo XX, aproximadamente el 60 % de la población de Perú era quechua-hablante, mientras que a comienzos del siglo XXI, sólo el 15 % lo era. No obstante, en términos absolutos, existen más hablantes de quechua hoy que en el siglo XIX.

El quechua se ha visto afectado por procesos de urbanización y desplazamiento lingüístico, pero también ha sido objeto de esfuerzos de revitalización. Un tercio de las personas que hablan quechua vive en entornos urbanos, lo que evidencia su adaptación a nuevas realidades sociales.

Previo a la llegada de los españoles, el idioma quechua era principalmente una lengua oral. La información se registraba a través de los quipus, un sistema de nudos utilizado por los incas para el almacenamiento de datos.

La escritura en quechua se desarrolló tras la colonización y ha sido objeto de diversas reformas. En 1985, el gobierno peruano oficializó un alfabeto con 34 grafías mediante la Resolución Ministerial N° 1218-85-ED. Posteriormente, se estableció el uso de un sistema de tres vocales (a, i, u) en 2013, lo que ha generado debates entre académicos y hablantes nativos.

El quechua ha sido utilizado en la educación intercultural bilingüe desde 1996 ([Hornberger and Coronel-Molina, 2004](#)). Sin embargo, su enseñanza enfrenta retos, como la falta de materiales didácticos y la estigmatización del idioma en ciertos sectores. A pesar de esto, en las últimas décadas ha habido un resurgimiento del quechua en medios de comunicación, música y producciones audiovisuales, contribuyendo a su revitalización.

Los avances en IA han abierto nuevas oportunidades para la preservación y enseñanza del quechua. Modelos de lenguaje como los LLMs pueden facilitar el aprendizaje de vocabulario, gramática y cultura a través de herramientas personalizadas ([Jafari,](#)

2023). Sin embargo, la escasez de datos en quechua plantea desafíos para la investigación en IA y PLN (Xiangjie et al., 2020).

El quechua es una lengua de gran valor histórico y cultural, pero enfrenta riesgos de desaparición en varias de sus variedades. A pesar de la disminución del número de hablantes, iniciativas de revitalización han permitido su continuidad y adaptación en la sociedad moderna. La incorporación de tecnologías como la IA podría contribuir significativamente a su preservación y difusión.

Tabla 2.1: *Condición de vitalidad de las variedades del quechua en el Perú*

Rama	Variedad	Departamento donde se habla	Estado de vitalidad	Corpus en este trabajo
Quechua amazónico	Kichwa amazónico: Pastaza, Napo, Putumayo, Tigre, Alto Napo, Santarrosino-Madre de Dios, Chachapoyas y San Martín	Loreto, Madre de Dios, Amazonas y San Martín	En peligro	—
Quechua nor-teño	Quechua Cajamarca	Cajamarca	Seramente en peligro	—
	Quechua Inkawasi Kañaris	Lambayeque y Piura	Vital	—
Quechua central	Quechua Pataz	La Libertad	Vital	—
	Quechua Cajatambo, Oyón, Huaura	Lima	Seramente en peligro	—
	Quechua Yauyos	Lima	Seramente en peligro	—
	Quechua Áncash	Áncash	Vital	—
	Quechua Huánuco	Huánuco	En peligro	—
	Quechua Pasco	Pasco	Seramente en peligro	—
	Quechua Wanka	Junín	Seramente en peligro	—
Quechua sureño	Quechua Chanka	Huancavelica, Ayacucho y Apurímac	Vital	Siminchik
	Quechua Collao	Apurímac, Cusco, Puno, Arequipa y Moquegua	Vital	Siminchik, qcc, Médico, LAQ-LA

Nota. Adaptado del Estado de Vitalidad del Quechua (Ministerio de Educación del Perú, 2018). La última columna indica en qué corpus de los empleados en este trabajo aparece cada variedad.

2.2. Morfología aglutinante del quechua

El quechua es una lengua de tipología aglutinante y marcadamente sufijante (Ortega et al., 2020). Esto significa que las palabras se construyen encadenando sufijos a una raíz, donde cada sufijo aporta una unidad de significado gramatical o semántico bien definida (número, caso, persona, tiempo, aspecto, evidencialidad, entre otros). En el sustantivo, el quechua Cusco-Collao expresa mediante sufijos de caso las relaciones que el español codifica a través de preposiciones (Cahuana Q., 2007). Como consecuencia, una sola palabra en quechua puede expresar lo que en español requiere una frase completa de varias palabras (Moreno Veliz, 2021).

La Tabla 2.2 ilustra este mecanismo a partir de la raíz nominal *wasi* (“casa”), mostrando cómo la adición de distintos sufijos de caso y de número modifica y precisa el significado.

Tabla 2.2: *Ejemplo de sufijación a partir de la raíz wasi (“casa”)*

Palabra en quechua	Segmentación	Traducción (sufijo)
wasi	wasi	casa
wasita	wasi-ta	a la casa (acusativo)
wasiman	wasi-man	hacia la casa (ilativo)
wasimanta	wasi-manta	desde la casa (ablativo)
wasipi	wasi-pi	en la casa (locativo)
wasiwan	wasi-wan	con la casa (instrumental)
wasipaq	wasi-paq	para la casa (dativo)
wasikuna	wasi-kuna	casas (plural)
wasipiñan	wasi-pi-ña-n	ya en la casa

Nota. Sufijos de caso y número del sustantivo en quechua Cusco-Collao, según Cahuana Q. (2007): *-ta* (acusativo), *-man* (ilativo), *-manta* (ablativo), *-pi* (locativo), *-wan* (instrumental), *-paq* (dativo) y *-kuna* (pluralizador). La última fila (*wasipiñan*, “ya en la casa”, ejemplo tomado de la misma fuente) ilustra el encadenamiento de varios sufijos sobre una misma raíz: locativo *-pi*, *-ña* (“ya”) y el validador *-n*.

Esta capacidad combinatoria tiene una consecuencia directa y crítica para el procesamiento computacional: la explosión del vocabulario. Dado que cada raíz puede generar numerosas formas superficiales distintas según la combinación de sufijos, el número de palabras únicas crece de manera mucho más acelerada que en lenguas de morfología más simple como el español. Este fenómeno puede cuantificarse mediante la razón tipo/token (proporción de palabras únicas respecto al total de palabras). Sobre el conjunto de prueba del dataset Médico empleado en este trabajo, el lado en quechua presenta una razón tipo/token de 0,76, frente a 0,44 del lado en español correspondiente; es decir, el quechua exhibe casi el doble de diversidad léxica sobre el mismo contenido semántico. Esta dispersión incrementa la tasa de palabras fuera de vocabulario (Palabras Fuera de Vocabulario (*Out-Of-Vocabulary*) (OOV)) durante la

inferencia y dificulta el aprendizaje de los modelos en escenarios de datos limitados (Ortega et al., 2020; Moreno Veliz, 2021).

2.2.1. Tokenización de subpalabras en el quechua

Para mitigar el problema de la explosión de vocabulario, los modelos de lenguaje modernos no operan directamente sobre palabras completas, sino sobre *subpalabras* (*subword units*). Algoritmos como Byte-Pair Encoding (BPE) (NLLB Team et al., 2022) y su variante SentencePiece empleada por el modelo NLLB utilizado en este trabajo construyen un vocabulario de fragmentos frecuentes a partir de los datos de entrenamiento, de modo que una palabra desconocida pueda representarse como una secuencia de fragmentos conocidos en lugar de un único token fuera de vocabulario.

Sin embargo, estos algoritmos determinan los fragmentos según su frecuencia estadística en el corpus, sin ningún conocimiento de la estructura morfológica de la lengua. En consecuencia, las fronteras de los subtokens no coinciden con las fronteras reales entre morfemas, y una lengua subrepresentada en los datos de entrenamiento como el quechua termina fragmentándose en un número mayor de piezas que una lengua bien representada. La Tabla 2.3 muestra la fragmentación de palabras reales del corpus médico bajo el tokenizador de NLLB.

Tabla 2.3: *Fragmentación de palabras quechua bajo el tokenizador de NLLB (SentencePiece)*

Palabra en quechua	N.º de subtokens	Fragmentación
nanaywanchkanki	4	nanay wan chkan ki
ruwananchikpuni	3	ruwanan chik puni
llocsimunñachu	4	llocs imun ñ achu
chinkaykuywanmi	5	ch ink ayku y wanmi
wiqsayakuranki	7	wi q say ak ur an ki

Nota. Fragmentación producida por el tokenizador de NLLB sobre palabras reales del corpus médico. Las divisiones responden a la frecuencia estadística de los fragmentos en los datos de entrenamiento y no a las fronteras morfológicas del quechua: mientras que en algunos casos un fragmento coincide con una raíz o sufijo reconocible (por ejemplo, *nanay* en la primera fila), en otros (como *wiqsayakuranki*) la palabra se descompone en piezas sin correspondencia lingüística.

Al medir esta fragmentación sobre la totalidad del conjunto de prueba del dataset Médico, el tokenizador de NLLB genera en promedio 2,31 subtokens por cada palabra en quechua, frente a 1,31 subtokens por palabra en español, es decir, una fragmentación 1,8 veces mayor sobre el mismo contenido. Este mismo patrón se reproduce con tokenizadores de otras arquitecturas: el tokenizador del modelo GPT-4o produce 2,92 subtokens por palabra en quechua frente a 1,62 en español, confirmando que se trata de una característica estructural de la lengua y no de un tokenizador particular.

Esta fragmentación más severa tiene implicaciones directas sobre el desempeño de los modelos. Al descomponerse en más piezas, las secuencias en quechua resultan más largas y sus dependencias más difíciles de modelar; además, un error en un único subtoken puede alterar la interpretación de toda la palabra. Este fenómeno contribuye a explicar los valores bajos de BLEU reportados en el Capítulo 5, así como la conveniencia de utilizar ChrF, que al operar a nivel de caracteres resulta más robusta ante estas variaciones morfológicas (Popović, 2015).

2.3. Lenguas de bajos recursos

En el campo del PLN y la traducción automática, el término *lengua de bajos recursos* (del inglés *low-resource language*) hace referencia a aquellos idiomas para los cuales existe una disponibilidad limitada de datos digitales anotados, corpus paralelos y herramientas computacionales de procesamiento (Besacier et al., 2014). Esta escasez condiciona directamente la capacidad de entrenar modelos de lenguaje con un rendimiento comparable al obtenido en lenguas de alto recurso, como el inglés, el chino mandarín o el español (Joshi et al., 2020).

El quechua constituye un caso representativo de lengua de bajos recursos en el contexto de la traducción automática. Los trabajos de investigación que abordan esta tarea reportan de manera recurrente los mismos obstáculos: la escasez de corpus paralelos alineados quechua-español, con colecciones de audio para traducción de voz que en algunos casos no superan las dos horas de duración (Ortega et al., 2025; Zafar et al., 2024); la morfología aglutinante del quechua, que genera vocabularios de gran tamaño y alta tasa de palabras fuera de vocabulario durante la inferencia (Moreno Veliz, 2021; Ortega et al., 2020); y la limitada representación del quechua en modelos preentrenados de gran escala, lo que obliga a recurrir a estrategias de *fine-tuning* con datos insuficientes (Cardenas et al., 2018; Ortega et al., 2025). Asimismo, se ha evidenciado que el aumento de datos mediante corpus sintéticos mejora el rendimiento solo cuando la calidad de los datos generados es adecuada, mientras que la incorporación de datos de baja calidad deteriora las métricas de evaluación (Ortega et al., 2025).

2.4. Aprendizaje Automático

El Aprendizaje Automático (*Machine Learning*) (ML) es una disciplina dentro de la IA que permite a los sistemas informáticos mejorar su rendimiento en tareas específicas sin necesidad de programación explícita. Su funcionamiento se basa en el uso de modelos matemáticos y algoritmos que analizan grandes volúmenes de datos para identificar patrones y realizar tareas como predicción, clasificación y toma de decisiones (Castillo, 2022).

Uno de los aspectos clave del aprendizaje automático es su capacidad de extraer

información útil a partir de datos previos. Para lograrlo, los modelos deben ser entrenados con una gran cantidad de información, lo que implica un proceso computacionalmente costoso. Sin embargo, gracias a los avances en la capacidad de procesamiento de los ordenadores, el entrenamiento y uso de estos modelos se ha vuelto más accesible y eficiente ([Castillo, 2022](#)).

En el ámbito del PLN, el aprendizaje automático juega un papel fundamental, ya que permite analizar grandes volúmenes de texto y extraer significado a partir del contexto. Algoritmos de ML son ampliamente utilizados en aplicaciones como traducción automática, análisis de sentimientos, reconocimiento de voz y generación de texto ([Castillo, 2022](#)).

2.5. Procesamiento de Lenguaje Natural

El campo del PLN se ocupa de un conjunto de enfoques computacionales derivados de teorías que permiten analizar y representar textos escritos en lenguajes humanos. Esta área, que se encuentra en la intersección de la IA y la lingüística computacional, busca comprender la interacción entre las máquinas y el lenguaje humano ([Moreira et al., 2021](#)). La principal finalidad del PLN es dotar a las computadoras de la capacidad de entender, interpretar y generar texto y discurso de forma similar a la que lo hacen las personas ([Moreira et al., 2021](#)).

Para lograrlo, se desarrollan algoritmos y modelos computacionales que permiten a las máquinas manipular y procesar el lenguaje humano en sus diversas formas, tanto hablado como escrito. Entre las diversas aplicaciones del procesamiento de lenguaje natural se incluyen la traducción automática, el análisis de sentimientos, la extracción de información, la generación automática de texto, el procesamiento de voz, el reconocimiento de voz, la síntesis de voz y la respuesta a preguntas, entre otras ([Vajjala et al., 2020](#)).

De este modo, el PLN abarca un amplio espectro de tareas que tienen un impacto significativo en diversas áreas tecnológicas y científicas ([Moreira et al., 2021](#)).

2.6. Traducción automática

La traducción automática tiene como objetivo principal tomar un texto en un idioma y traducirlo a otro manteniendo su significado ([Hernández and Gómez, 2013](#)). Este proceso generalmente sigue tres etapas: primero, el texto original se convierte en una representación intermedia; luego, se ajusta esta representación según la morfología del idioma de destino; y finalmente, se transforma al lenguaje destino. Evaluar la calidad de la traducción es complejo, ya que la noción de "significado" no tiene una definición exacta y puede variar entre evaluadores. Sin embargo, con un enfoque estadístico y suficientes muestras, es posible obtener evaluaciones válidas y precisas. El alineamiento

de palabras, basado en corpus con traducciones de palabras en diferentes contextos, es fundamental en los modelos estadísticos actuales de traducción ([Hernández and Gómez, 2013](#)). A pesar de los avances, existen aún numerosos retos, especialmente cuando se trabaja con idiomas que tienen una estructura morfológica compleja o un orden de palabras distinto ([Hernández and Gómez, 2013](#)).

2.7. Traducción de voz a texto

La traducción de voz a texto es el proceso de convertir una señal de habla en un idioma origen en texto escrito en un idioma destino. Este problema combina dos áreas tradicionales del procesamiento del lenguaje natural y del habla: el ASR y la MT ([Besacier et al., 2014](#)). En la actualidad, existen dos enfoques principales para abordar esta tarea: el enfoque en cascada y el enfoque de traducción directa o *end-to-end* ([Kano et al., 2021](#)).

2.7.1. Enfoque en cascada

El método en cascada ha sido históricamente el más utilizado debido a que aprovecha tecnologías ya maduras y desarrolladas por separado ([Ortega et al., 2025](#)). En este enfoque, el sistema se divide en dos etapas consecutivas. En primer lugar, un modelo de reconocimiento automático del habla transcribe la señal de audio del idioma origen a texto. Posteriormente, un sistema de traducción automática recibe dicha transcripción como entrada y genera el texto traducido en el idioma destino.

De forma simplificada, el proceso puede representarse como:

Audio en idioma origen $\rightarrow$ Transcripción $\rightarrow$ Texto traducido

La principal ventaja de este enfoque es su modularidad, ya que cada componente puede entrenarse, evaluarse y mejorarse de manera independiente ([Koehn, 2010](#)). Además, permite reutilizar modelos de ASR y MT ya existentes, lo que reduce costos de desarrollo y facilita su implementación ([Zafar et al., 2024](#)).

Sin embargo, también presenta limitaciones importantes. La más relevante es la propagación de errores: si el sistema de reconocimiento de voz produce una transcripción incorrecta, esos errores pasan directamente al módulo de traducción y pueden deteriorar la calidad final ([Kano et al., 2021](#)). Asimismo, al separar ambas tareas, el sistema no aprovecha completamente la información acústica, prosódica o contextual presente en la señal de voz original.

2.7.2. Enfoque de traducción directa

Como alternativa al enfoque tradicional, investigaciones recientes han propuesto sistemas de traducción directa de voz a texto, también conocidos como modelos *end-to-end* (Kano et al., 2021; Ao et al., 2022). En este caso, un único modelo aprende a transformar directamente la señal de audio en el idioma origen al texto traducido en el idioma destino, sin necesidad de generar una transcripción intermedia.

Este proceso puede representarse como:

Audio en idioma origen $\rightarrow$ Texto traducido

Gracias al uso de arquitecturas neuronales profundas, como modelos basados en codificadores y decodificadores con mecanismos de atención o transformadores, estos sistemas pueden aprender relaciones conjuntas entre el contenido acústico y la traducción. Estudios como el de Kano et al. (Kano et al., 2021) muestran que es posible realizar traducción directa eliminando la necesidad de construir modelos separados para reconocimiento del habla y traducción.

Entre sus ventajas se encuentra la reducción de la propagación de errores entre módulos, una arquitectura más compacta y la posibilidad de aprovechar información presente en la señal de voz que podría perderse durante la transcripción (Ortega et al., 2025). No obstante, estos sistemas suelen requerir grandes volúmenes de datos paralelos de voz y traducción para su entrenamiento, recursos que no siempre están disponibles, especialmente para idiomas de bajos recursos (Besacier et al., 2014).

2.7.3. Comparación general

En términos generales, los sistemas en cascada continúan siendo ampliamente utilizados por su robustez y facilidad de implementación, especialmente cuando existen buenos modelos de reconocimiento y traducción para los idiomas involucrados (Zafar et al., 2024). Por otro lado, los modelos directos representan una línea de investigación prometedora, ya que buscan simplificar la arquitectura y mejorar la calidad final al aprender la tarea completa de forma conjunta (Ortega et al., 2025).

La elección entre ambos enfoques depende de factores como la disponibilidad de datos, los recursos computacionales, la calidad requerida y los idiomas considerados (Ortega et al., 2025; Zafar et al., 2024).

2.7.4. Formalización probabilística de los dos enfoques

Ambos enfoques pueden expresarse dentro de un marco probabilístico común, lo que permite precisar en qué difieren (Sperber and Paulik, 2020). La traducción de voz

consiste en hallar la secuencia de texto destino y más probable dada una señal de audio de entrada x :

$$\hat{y} = \arg \max_y P(y | x). \quad (2.1)$$

Introduciendo la transcripción intermedia t en el idioma origen como variable latente, la probabilidad del texto destino puede descomponerse de forma exacta marginalizando sobre todas las transcripciones posibles:

$$P(y | x) = \sum_t P(y | t) P(t | x), \quad (2.2)$$

donde $P(t | x)$ corresponde al modelo de reconocimiento de voz (ASR) y $P(y | t)$ al modelo de traducción (MT). La Ecuación 2.2 expresa el sistema ideal, que considera la incertidumbre de todas las transcripciones posibles ponderadas por su probabilidad.

El enfoque en cascada realiza una aproximación de esta expresión: en lugar de sumar sobre todas las transcripciones, selecciona únicamente la más probable, $\hat{t} = \arg \max_t P(t | x)$, y traduce sobre ella:

$$P(y | x) \approx P(y | \hat{t}). \quad (2.3)$$

Esta simplificación es la que da lugar a la propagación de errores. Al condicionar la traducción sobre una única transcripción $\hat{t}$, el sistema descarta la incertidumbre del reconocimiento de voz: si $\hat{t}$ contiene errores, estos se transmiten al módulo de traducción como si fueran entrada correcta, y la información acústica que podría haber ayudado a corregirlos se pierde (Kano et al., 2021). El costo de la aproximación de la Ecuación 2.3 respecto a la marginalización exacta de la Ecuación 2.2 es, precisamente, el error propagado desde el ASR.

El enfoque de traducción directa, en cambio, modela $P(y | x)$ mediante un único modelo, sin construir la variable intermedia t ni realizar la aproximación anterior. En teoría, esto le permite aprovechar toda la información de la señal acústica y evitar la propagación de errores, por lo que debería resultar superior. En la práctica, sin embargo, estimar $P(y | x)$ de forma directa exige aprender conjuntamente la correspondencia acústica y semántica a partir de pares audio-traducción, lo que requiere volúmenes de datos considerablemente mayores que los necesarios para entrenar por separado cada componente del enfoque en cascada (Ortega et al., 2025; Besacier et al., 2014). Esta tensión, es decir, la ventaja teórica del enfoque directo frente a su mayor demanda de datos, es la que se examina empíricamente en el Capítulo 5 bajo dos regímenes de disponibilidad de datos.

2.7.5. Alucinación en modelos de traducción de voz

Un fenómeno particularmente relevante para el enfoque de traducción directa es la *alucinación*: la generación de una salida fluida y gramaticalmente correcta en el idioma destino, pero semánticamente desconectada de la entrada (Raunak et al., 2021). A diferencia de un error de traducción ordinario, en el que la salida conserva alguna relación con el contenido original, una alucinación produce texto plausible que no corresponde a lo que expresó el hablante, lo que la hace especialmente peligrosa en dominios sensibles como el clínico.

La literatura atribuye este comportamiento, en parte, al *sesgo de exposición* (*exposure bias*) del entrenamiento secuencia a secuencia: durante el entrenamiento el modelo siempre recibe el token de referencia previo, mientras que en inferencia debe condicionar sobre sus propias predicciones, lo que puede llevarlo a divergir hacia secuencias fluidas pero infundadas (Wang and Sennrich, 2020). Este efecto se agrava bajo *cam-bio de dominio* (*domain shift*), es decir, cuando los datos de prueba difieren de los de entrenamiento (Wang and Sennrich, 2020), condición que se cumple en este trabajo, donde el corpus de prueba pertenece al dominio médico.

El régimen de datos escasos acentúa el problema. Se ha documentado que la calidad de la traducción neuronal se degrada de forma abrupta cuando el volumen de datos de entrenamiento es reducido (Koehn and Knowles, 2017), situación característica de las lenguas de bajos recursos y, en particular, del corpus de audio médico empleado en este trabajo. En estas condiciones, un modelo directo con capacidad insuficiente de anclaje a la entrada tiende a compensar generando lenguaje fluido a costa de la fidelidad semántica. Este marco teórico permite interpretar el comportamiento del modelo de traducción directa observado en el Capítulo 5, donde produce salidas gramaticalmente correctas pero desconectadas del audio de origen.

2.8. Modelo de lenguaje

En el ámbito del procesamiento de lenguaje natural y la inteligencia artificial, los modelos de lenguaje desempeñan un papel crucial al permitir la predicción y generación de texto en función del contexto y las secuencias de palabras. Básicamente, se trata de modelos estadísticos o basados en aprendizaje automático que identifican patrones y estructuras presentes en el lenguaje humano (Kłosowski, 2018). Estos modelos son entrenados utilizando grandes volúmenes de datos textuales, lo que les permite aprender las probabilidades de aparición de palabras o secuencias en diferentes situaciones lingüísticas. A medida que avanzan en su entrenamiento, los modelos logran captar la gramática, el significado y las relaciones entre las palabras dentro de un contexto natural (Bengio et al., 2000).

Los modelos de lenguaje se aplican en diversas áreas, tales como la traducción automática, la corrección gramatical, la búsqueda de información, la generación auto-

mática de texto, la realización de resúmenes y el análisis de sentimientos, entre muchas otras (Moreira et al., 2021). Algunos de los modelos más populares incluyen el Generative Pre-trained Transformer (GPT) y Bidirectional Encoder Representations from Transformers (BERT), que han mostrado un rendimiento excepcional en tareas relacionadas con el procesamiento de lenguaje natural (Radford et al., 2018). Estos avanzados modelos de lenguaje aprovechan redes neuronales y mecanismos de atención para profundizar su comprensión del lenguaje humano.

2.8.1. Large Language Models (LLMs)

Los Grandes Modelos de Lenguaje (LLMs, por sus siglas en inglés) están basados en arquitecturas de tipo Transformer, diseñados para analizar y producir texto de forma coherente y contextual. Estos modelos se entrenan con enormes cantidades de datos y utilizan técnicas como el aprendizaje profundo, el autoaprendizaje y el ajuste fino para responder preguntas, traducir idiomas, resumir textos y realizar una amplia variedad de tareas de PLN (Radford et al., 2018).

2.8.2. Embedding

Dentro del ámbito del procesamiento de lenguaje natural, un embedding (incrustación o representación vectorial) se refiere a la técnica de convertir palabras, frases o documentos en vectores numéricos de alta dimensionalidad (Bhoir et al., 2017). El término embedding proviene de su significado matemático, que se refiere a la asignación de una estructura o espacio a otro, se utiliza para asignar palabras a un espacio multidimensional donde sus relaciones semánticas y contextuales se representan de manera matemática, esto permite que las palabras se manejen y comparen de manera más efectiva en aplicaciones de procesamiento de lenguaje (Jurafsky and Martin, 2019). Estos vectores se utilizan para representar e identificar las conexiones semánticas y sintácticas entre las palabras en un espacio vectorial continuo (Bhoir et al., 2017). Los embeddings son fundamentales en PLN porque al permitir que las computadoras entiendan y manejen el lenguaje humano de forma más eficiente y significativa (Mikolov et al., 2013).

En la práctica, un modelo neuronal no opera sobre los tokens como símbolos discretos, sino sobre los vectores asociados a ellos. Cada token del vocabulario ocupa una fila de una matriz de embeddings de dimensión $|V| \times d$, donde $|V|$ corresponde al tamaño del vocabulario y d a la dimensión del espacio de representación; la primera operación que realiza cualquiera de estos modelos consiste en sustituir cada token de entrada por su fila correspondiente (Jurafsky and Martin, 2019). Esta distinción resulta relevante para el presente trabajo porque, tal como se describe en la Sección 2.2.1, las palabras del quechua se fragmentan en un número elevado de subtokens cuyas fronteras no coinciden con las de sus morfemas. En consecuencia, una palabra como *wiqsayakuranki* no dispone de un vector propio dentro del modelo: su representación debe reconstruirse componiendo los vectores de siete fragmentos que carecen de significado morfológico

individual. La subrepresentación de una lengua de bajos recursos se manifiesta así no solo como escasez de datos, sino como una calidad deficiente del espacio vectorial en el que dicha lengua queda representada (Joshi et al., 2020).

Históricamente, los primeros embeddings fueron de tipo estático: cada palabra recibía un único vector con independencia del contexto en el que apareciera, como ocurre en el modelo Word2Vec (Mikolov et al., 2013). Las arquitecturas de tipo Transformer introdujeron los embeddings contextuales, en los que el vector asignado a un token se recalcula en cada capa en función de los demás elementos de la secuencia mediante los mecanismos de atención (Vaswani et al., 2017). Los tres modelos empleados en este trabajo, SpeechT5, Whisper y NLLB, pertenecen a esta segunda familia y comparten además una condición metodológica relevante: sus matrices de embeddings no se construyen desde cero, sino que se heredan del preentrenamiento sobre grandes volúmenes de datos y se reajustan durante el ajuste fino descrito en la Sección 2.9. El detalle de qué magnitud se convierte en vector dentro de cada arquitectura y qué ocurre con esas representaciones durante los experimentos realizados se presenta en la Sección 4.5.4.

2.8.3. Tokens

Dentro del ámbito del procesamiento de lenguaje natural y la informática, un “token” se refiere a la unidad más pequeña de un texto o secuencia que se utiliza como unidad de procesamiento. Estos tokens pueden ser palabras individuales, sub-palabras (como prefijos o sufijos), caracteres o incluso partes más pequeñas de una palabra, dependiendo de cómo se defina el tokenizador (Jurafsky and Martin, 2019). La forma en que se definen estas unidades es especialmente relevante para lenguas de morfología aglutinante como el quechua, cuyo efecto sobre la tokenización de subpalabras se analiza en la Sección 2.2.1. Los tokens son la base para que las computadoras comprendan y procesen el lenguaje humano en tareas como la traducción automática, análisis de sentimientos y la generación de texto, ya que permiten descomponer y trabajar con precisión en el texto de entrada (Radford et al., 2018).

2.8.4. Arquitectura Transformer

La arquitectura Transformer se destaca por no depender de la recurrencia, sino que emplea un mecanismo de atención para identificar las relaciones entre los elementos de entrada y salida de manera global (Vaswani et al., 2017). Esta característica le permite procesar secuencias de datos en paralelo, lo que resulta en una mayor eficiencia, y facilita la captura de relaciones a largo plazo en los datos. En el proceso de trabajo de un Transformer, las secuencias de entrada se fragmentan en *tokens*, y mediante el mecanismo de atención el modelo determina la importancia de cada token en relación con los demás dentro de la secuencia.

Gracias a la atención contextual, Transformer puede entender de manera efectiva la estructura y las dependencias de los datos o textos, lo que lo hace sumamente efi-

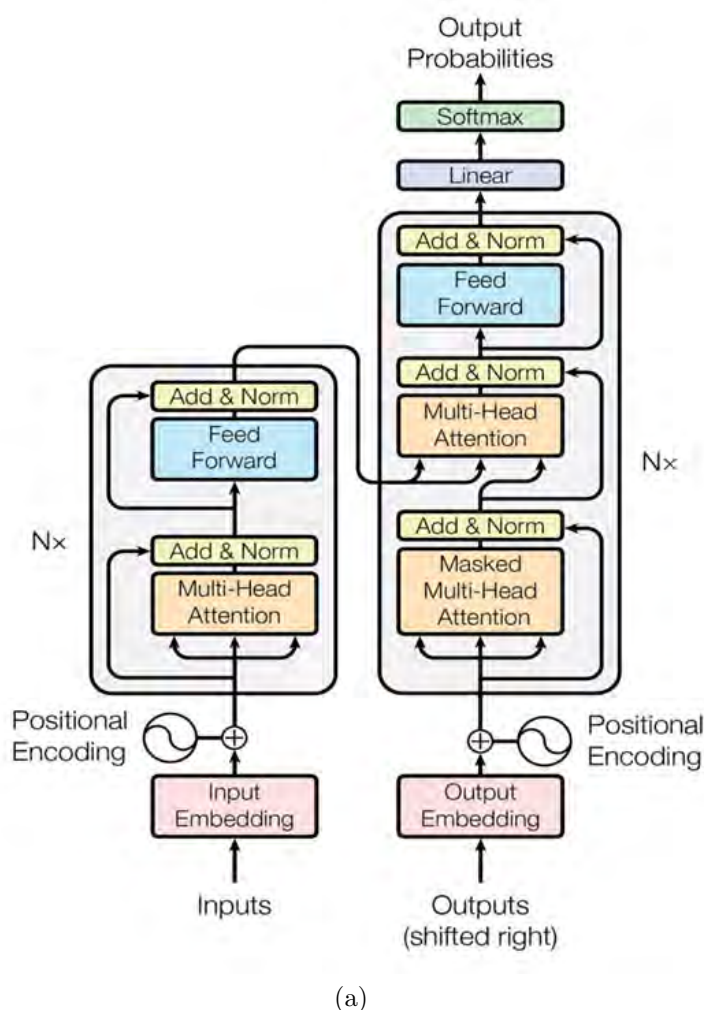


Figura 2.1: Arquitectura del modelo Transformer. La imagen ilustra los componentes clave, incluyendo la autoatención, las capas de codificación y decodificación, y los mecanismos de normalización y proyección (Vaswani et al., 2017).

ciente en tareas como la traducción automática, la generación de texto y el resumen de contenido. Un ejemplo destacado de esta arquitectura es el modelo GPT (Generative Pre-trained Transformer), que ha demostrado un rendimiento excepcional en múltiples aplicaciones de procesamiento de lenguaje natural (Radford et al., 2018). Esta innovación ha transformado el campo, superando las limitaciones de modelos anteriores y estableciendo un nuevo estándar en el procesamiento de lenguaje y otras áreas relacionadas con secuencias de datos (Vaswani et al., 2017).

El componente central de la arquitectura es el mecanismo de *atención de producto punto escalado* (*scaled dot-product attention*), base de los tres modelos empleados en este trabajo. Dadas tres matrices: consultas Q , claves K y valores V , la atención se define como (Vaswani et al., 2017):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \quad (2.4)$$

donde d_k es la dimensión de las claves y el factor $\sqrt{d_k}$ estabiliza los gradientes al evitar que los productos punto crezcan en magnitud con la dimensión. La función *softmax* convierte las similitudes entre consultas y claves en pesos de atención que ponderan los valores, de modo que cada elemento de la secuencia se representa en función de su relación con todos los demás.

Para capturar distintos tipos de relaciones de forma simultánea, el Transformer aplica varias proyecciones de atención en paralelo, mecanismo denominado *atención multi-cabeza* (*multi-head attention*) (Vaswani et al., 2017):

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O, \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \quad (2.5)$$

donde cada cabeza head_i emplea matrices de proyección propias W_i^Q , W_i^K y W_i^V , y W^O es la proyección de salida que combina las h cabezas. Esta capacidad de atender a la secuencia completa en paralelo, sin recurrir a la recurrencia, es la que permite a los modelos basados en Transformer modelar dependencias de largo alcance; relevantes, por ejemplo, para relacionar sufijos distantes dentro de una misma palabra aglutinante, y procesar la señal de entrada de forma eficiente.

2.8.5. SpeechT5

SpeechT5 (Ao et al., 2022) es un modelo multimodal basado en la arquitectura Transformer de tipo encoder-decoder (Vaswani et al., 2017), diseñado para unificar tareas de procesamiento de voz y texto, tales como ASR, síntesis de voz (Texto a Voz (*Text To Speech*) (TTS)) y Traducción de Voz (*Speech Translation*) (ST). Su principal fortaleza radica en el aprendizaje conjunto de representaciones acústicas y lingüísticas dentro de un mismo espacio latente.

El codificador (encoder) es responsable de procesar tanto señales de voz como texto, los cuales son previamente transformados mediante módulos de preprocesamiento (prenet). En el caso del audio, el *Speech Encoder Prenet* convierte características acústicas, como espectrogramas, en representaciones vectoriales, mientras que el *Text Encoder Prenet* transforma los tokens de texto en embeddings (Ao et al., 2022). Posteriormente, el encoder proyecta ambas modalidades en un espacio latente compartido, permitiendo que información lingüística y acústica sea representada de manera unificada.

El decodificador (decoder), también basado en Transformer (Vaswani et al., 2017), toma como entrada estas representaciones latentes y genera salidas en función de la

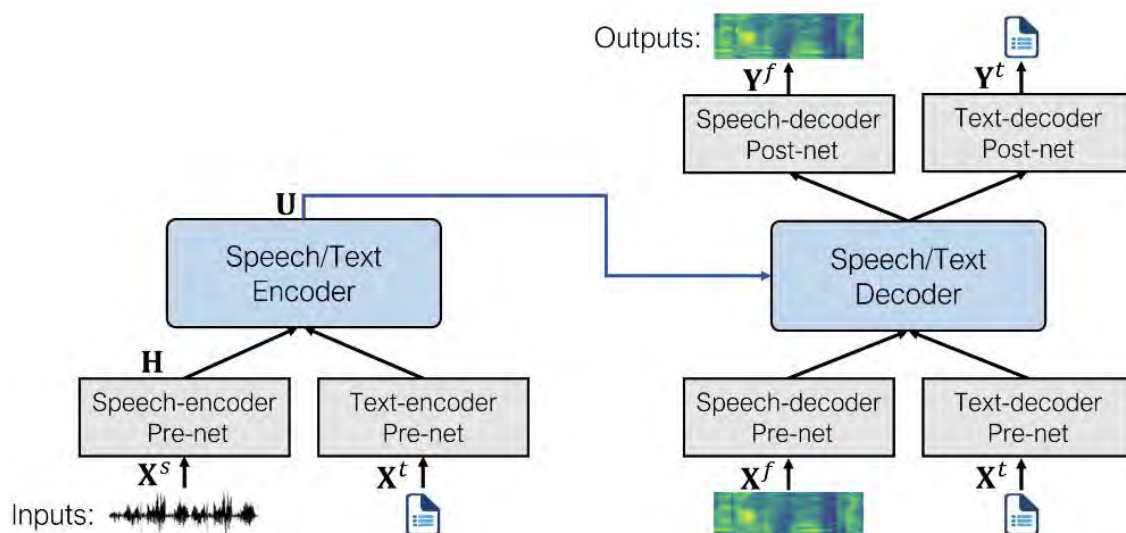


Figura 2.2: Arquitectura del modelo SpeechT5.

tarea objetivo. Para ello, emplea módulos de postprocesamiento (*postnet*) específicos para cada modalidad. El *Text Decoder Postnet* se encarga de generar secuencias de texto, lo cual resulta útil en tareas como ASR o traducción (Ao et al., 2022). Por otro lado, el *Speech Decoder Postnet* produce representaciones acústicas, típicamente en forma de espectrogramas Mel, que posteriormente pueden ser convertidos en señal de audio mediante un vocoder en aplicaciones de TTS (Ao et al., 2022).

Un aspecto fundamental de SpeechT5 es el uso de un espacio latente compartido (Ao et al., 2022), el cual constituye el núcleo del modelo. Tanto el texto como el audio son codificados en este espacio común mediante el encoder, lo que permite establecer una correspondencia directa entre ambas modalidades (Ao et al., 2022; Vaswani et al., 2017). Esta característica habilita la realización de múltiples tareas dentro de un mismo modelo, incluyendo la conversión de texto a voz (TTS), voz a texto (ASR), voz a voz (por ejemplo, conversión de voz) y texto a texto (tareas de procesamiento de lenguaje natural) (Ao et al., 2022). En consecuencia, SpeechT5 logra unificar distintos problemas bajo un mismo marco arquitectónico, aprovechando el aprendizaje conjunto de representaciones multimodales (Ao et al., 2022).

Cabe precisar que este espacio latente compartido constituye, en los términos de la Sección 2.8.2, un espacio de embeddings multimodal: tanto los fragmentos de audio como los tokens de texto se convierten en vectores de una misma dimensión y coexisten en una única geometría. De ello se desprende que la calidad de la traducción directa depende de que las representaciones del audio en quechua y las del texto en español queden efectivamente próximas dentro de dicho espacio, alineamiento que el modelo debe aprender a partir de los pares de entrenamiento disponibles.

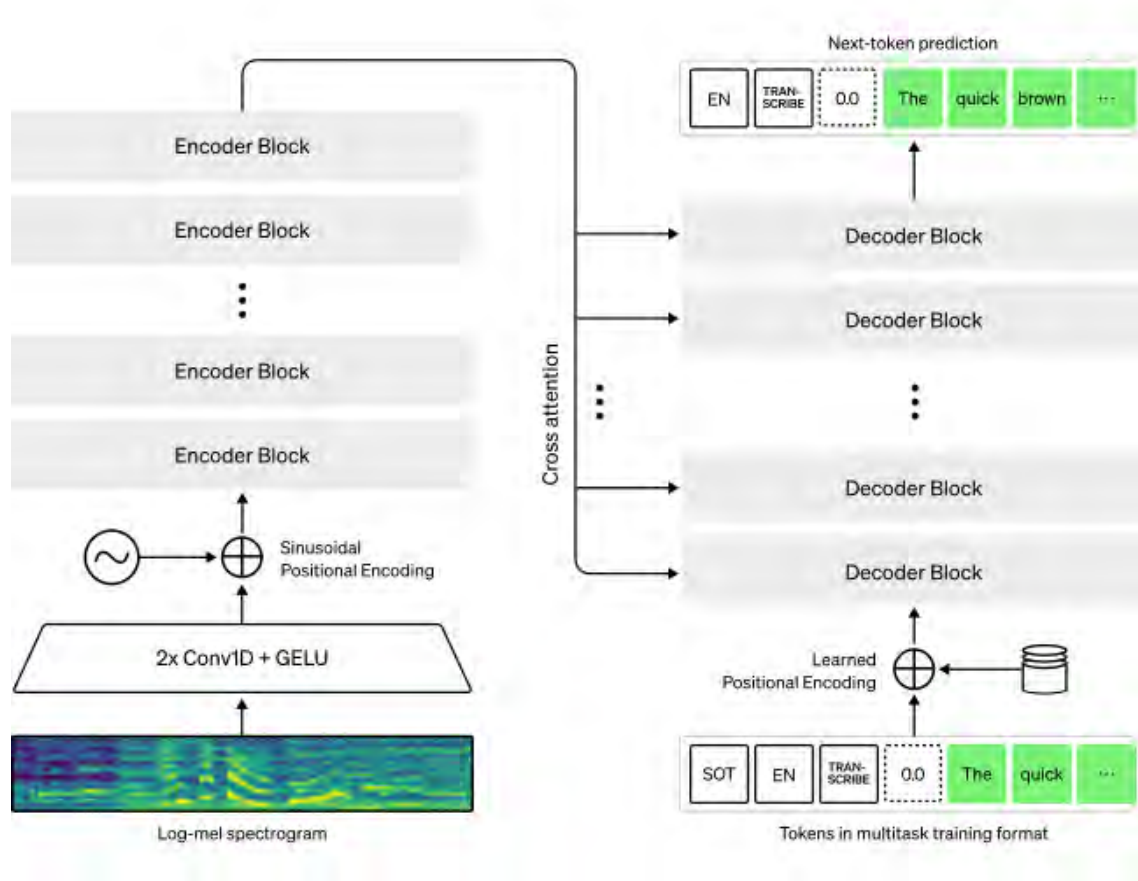


Figura 2.3: Arquitectura del modelo Whisper.

2.8.6. Whisper

El modelo Whisper (Radford et al., 2023) es un sistema de ASR basado en una arquitectura Transformer encoder-decoder (Vaswani et al., 2017), entrenado con aproximadamente 680,000 horas de audio multilingüe (Radford et al., 2023). Su diseño multitarea le permite realizar transcripción, traducción y detección de idioma dentro de un mismo modelo (Radford et al., 2023).

Como entrada, Whisper utiliza espectrogramas log-Mel, los cuales son procesados por un encoder bidireccional que captura patrones temporales en la señal de audio (Radford et al., 2023). El decoder autorregresivo genera la secuencia de salida, condicionado por la representación acústica y tokens especiales que indican idioma y tarea (Radford et al., 2023; Vaswani et al., 2017).

En términos de representación vectorial, el espectrograma log-Mel es proyectado por las capas convolucionales de entrada a una secuencia de embeddings acústicos, sobre los cuales opera el encoder (Radford et al., 2023). Los tokens especiales que indican el idioma y la tarea no constituyen una excepción: también ingresan al decoder como vectores de la matriz de embeddings del modelo, de modo que la selección del idioma se realiza condicionando la generación mediante una representación aprendida

y no mediante una regla externa.

El modelo emplea mecanismos de *self-attention* y *cross-attention* (Vaswani et al., 2017), y durante la inferencia puede utilizar estrategias como *greedy decoding* o *beam search* (Goodfellow et al., 2016). Gracias a su entrenamiento a gran escala, Whisper presenta alta robustez frente a ruido y variabilidad acústica (Radford et al., 2023), aunque su desempeño puede verse limitado en lenguas de bajos recursos como el quechua (Besacier et al., 2014; Joshi et al., 2020).

2.8.7. NLLB

El modelo NLLB (NLLB Team et al., 2022) es un sistema de MT multilingüe basado en Transformers (Vaswani et al., 2017), diseñado para soportar traducción directa entre más de 200 idiomas (NLLB Team et al., 2022).

El encoder transforma la secuencia de entrada en una representación semántica compartida, mientras que el decoder genera la traducción de forma autorregresiva utilizando mecanismos de atención (Vaswani et al., 2017; NLLB Team et al., 2022). El modelo emplea tokenización basada en subpalabras mediante SentencePiece (NLLB Team et al., 2022), lo que permite manejar vocabularios amplios y mejorar la generalización.

La tokenización mediante SentencePiece indexa una matriz de embeddings de subpalabras compartida por los más de doscientos idiomas que el modelo admite, a la que se añade un token de idioma (`quy_Latn` para el quechua y `spa_Latn` para el español) que actúa como vector de condicionamiento y determina la lengua de salida (NLLB Team et al., 2022). Dado que las filas de esa matriz asociadas a fragmentos propios del quechua se actualizaron durante el preentrenamiento con un volumen de ejemplos muy inferior al de las lenguas de alto recurso, su representación resulta menos precisa, lo que refuerza la necesidad del ajuste fino aplicado en este trabajo (Joshi et al., 2020).

NLLB fue entrenado sobre grandes volúmenes de datos paralelos obtenidos mediante técnicas de minería de texto multilingüe (NLLB Team et al., 2022), combinadas con procesos de filtrado y balanceo para reducir sesgos hacia idiomas de alto recurso. A pesar de su gran escala (con decenas de miles de millones de parámetros), su rendimiento puede verse afectado en idiomas altamente subrepresentados (Joshi et al., 2020), lo que justifica la necesidad de adaptación mediante fine-tuning en datasets específicos como los utilizados en este trabajo (Goodfellow et al., 2016; Ortega et al., 2025).

2.9. Aprendizaje por transferencia y adaptación de dominio

El *aprendizaje por transferencia* (*transfer learning*) consiste en reutilizar el conocimiento adquirido por un modelo durante un preentrenamiento sobre grandes volúmenes de datos generales para, posteriormente, especializarlo en una tarea o dominio con datos más limitados (Goodfellow et al., 2016). En los modelos empleados en este trabajo; SpeechT5, Whisper y NLLB, el preentrenamiento provee representaciones lingüísticas y acústicas de propósito general, mientras que la especialización se realiza mediante *ajuste fino* (*fine-tuning*): un reentrenamiento sobre datos del dominio objetivo que ajusta los pesos preexistentes en lugar de partir de una inicialización aleatoria.

El fine-tuning se distingue del preentrenamiento en su punto de partida y en su objetivo: el preentrenamiento aprende representaciones amplias a partir de datos heterogéneos, mientras que el fine-tuning las adapta a una distribución específica empleando un número mucho menor de ejemplos. Esta distinción es central en el presente trabajo, donde la diferencia entre la Configuración 1 (entrenamiento sin datos médicos) y la Configuración 2 (con incorporación del corpus médico) constituye, precisamente, un experimento de adaptación de dominio: se mide en qué grado la especialización sobre datos clínicos mejora el desempeño respecto a un modelo entrenado únicamente con lenguaje general.

La adaptación de dominio en traducción automática ha sido estudiada de forma sistemática. El trabajo de Chu et al. (2017) compara empíricamente distintas estrategias; desde el ajuste fino directo hasta métodos de mezcla y ponderación de datos, y muestra que la eficacia de cada una depende de la cantidad de datos del dominio objetivo y de su distancia respecto al dominio general. Estos resultados enmarcan teóricamente la comparación entre configuraciones realizada en el Capítulo 5 y respaldan la expectativa de que la incorporación de datos del dominio médico, aun en cantidades reducidas, produzca mejoras medibles sobre el desempeño en el dominio general.

2.10. Métricas de evaluación

La evaluación automática de sistemas de traducción busca estimar qué tan cercana es una traducción generada por un modelo respecto a una o más traducciones de referencia elaboradas por humanos (Koehn, 2010). En traducción automática, estas métricas permiten comparar modelos de manera objetiva, reproducible y a gran escala (Papineni et al., 2002), aunque no reemplazan completamente la evaluación humana, la cual considera aspectos como fluidez, adecuación semántica y naturalidad del lenguaje (Koehn, 2010).

Para este trabajo se emplearon las métricas BLEU y ChrF, debido a que ambas capturan dimensiones complementarias de la calidad de traducción (Hendy et al., 2023). Mientras BLEU se enfoca en coincidencias de palabras o secuencias de palabras

(Papineni et al., 2002), ChrF evalúa similitudes a nivel de caracteres (Popović, 2015), lo que resulta especialmente útil en lenguas con alta variación morfológica (Joshi et al., 2020).

2.10.1. BLEU

BLEU es una de las métricas más utilizadas para la evaluación de traducción automática (Papineni et al., 2002). Su objetivo es medir la similitud entre una traducción generada automáticamente y una o varias traducciones de referencia humanas mediante el cálculo de coincidencias de n-gramas (Papineni et al., 2002).

Formalmente, BLEU se define como:

$$\text{BLEU} = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (2.6)$$

donde p_n representa la precisión de n-gramas de orden n , w_n son pesos asociados (generalmente uniformes), y BP es el factor de penalización por brevedad. La precisión de n-gramas se calcula utilizando conteos recortados para evitar sobreestimaciones:

$$p_n = \frac{\sum \min(\text{count}_{cand}, \text{count}_{ref})}{\sum \text{count}_{cand}} \quad (2.7)$$

El factor de penalización por brevedad se define como:

$$BP = \begin{cases} 1 & \text{si } c > r \\ e^{(1-r/c)} & \text{si } c \leq r \end{cases} \quad (2.8)$$

donde c es la longitud de la traducción generada y r la longitud de la referencia.

El valor de BLEU está naturalmente acotado en el intervalo $[0, 1]$, lo que implica que se trata de una métrica normalizada. No obstante, en la práctica es común reportar su valor en una escala de 0 a 100, multiplicando el resultado por 100, lo que facilita su interpretación y comparación en la literatura.

BLEU ha sido ampliamente adoptada por su simplicidad, rapidez de cálculo y utilidad para comparar modelos en condiciones similares (Papineni et al., 2002). Además de traducción automática, también ha sido utilizada en tareas como resumen automático y generación de texto (Goodfellow et al., 2016).

Sin embargo, presenta limitaciones importantes. Al basarse en coincidencias superficiales de palabras, no captura adecuadamente la equivalencia semántica, la fluidez gramatical ni la adecuación contextual (Papineni et al., 2002).

A estas limitaciones se suma su inestabilidad estadística en conjuntos de prueba pequeños. Dado que BLEU se calcula sobre coincidencias de n-gramas a nivel de corpus, con pocas oraciones de referencia, como las 65 del *split* de prueba del dataset laqla, la métrica se vuelve sensible a variaciones mínimas y sus valores puntuales pueden no ser distinguibles de la variabilidad del muestreo (Koehn, 2010). Por ello, en conjuntos reducidos se recomienda acompañar el valor de BLEU con pruebas de significancia estadística, tal como se realiza en el Capítulo 5. Asimismo, el uso de una única traducción de referencia; habitual en lenguas de bajos recursos, penaliza traducciones correctas que difieren léxicamente de la referencia, subestimando la calidad real del sistema (Papineni et al., 2002).

Es importante, además, contextualizar las escalas de interpretación de uso común en la industria. Referencias frecuentes, como la que califica un valor de BLEU inferior a 10 como de utilidad práctica muy baja (Google Cloud, 2026), están calibradas para pares de idiomas de alto recurso evaluados con múltiples referencias, condiciones muy distintas a las de este trabajo. En un escenario de bajos recursos, con una única referencia y una lengua aglutinante subrepresentada, dichos umbrales no son directamente transferibles; en tales condiciones, una métrica orientada a caracteres como ChrF resulta más informativa que el valor absoluto de BLEU (Popović, 2015; Joshi et al., 2020).

2.10.2. ChrF

ChrF es una métrica basada en caracteres diseñada para evaluar traducciones automáticas mediante la comparación de n-gramas de caracteres entre la salida generada y la traducción de referencia (Popović, 2015).

A diferencia de BLEU, ChrF combina precisión y exhaustividad (*precision* y *recall*) mediante una media armónica ponderada. Su formulación es la siguiente:

$$\text{chrF}_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 P + R} \quad (2.9)$$

donde P representa la precisión de n-gramas de caracteres y R el recall, definidos como:

$$P = \frac{\text{n-gramas correctos}}{\text{n-gramas en la predicción}}, \quad R = \frac{\text{n-gramas correctos}}{\text{n-gramas en la referencia}} \quad (2.10)$$

El parámetro β controla la importancia relativa del recall, siendo común el uso de $\beta = 2$ (Popović, 2015).

Al igual que BLEU, ChrF produce valores en el intervalo $[0, 1]$, por lo que también se considera una métrica normalizada. En la práctica, estos valores suelen escalarse al

rango $[0, 100]$ para facilitar su interpretación.

Al operar a nivel de caracteres, ChrF resulta más robusta frente a variaciones ortográficas, flexiones gramaticales y diferencias morfológicas (Popović, 2015).

2.10.3. WER y CER

A diferencia de BLEU y ChrF, que evalúan la calidad de una traducción, la WER y la CER evalúan la calidad de una transcripción respecto a una referencia, por lo que constituyen las métricas estándar para sistemas de ASR (Jurafsky and Martin, 2019). En el presente trabajo se emplean para medir el desempeño del componente de reconocimiento de voz dentro del enfoque en cascada, de manera independiente a la calidad de la traducción final.

La WER se basa en la distancia de edición (o distancia de Levenshtein) entre la secuencia de palabras predicha y la de referencia, y se define como:

$$\text{WER} = \frac{S + D + I}{N} \quad (2.11)$$

donde S es el número de sustituciones, D el de eliminaciones (*deletions*), I el de inserciones y N el número total de palabras en la referencia. El valor resultante representa la proporción de palabras que deben modificarse para transformar la hipótesis en la referencia; un valor de 0 indica una transcripción perfecta, mientras que valores cercanos o superiores a 1 (100 %) indican transcripciones altamente erróneas.

La CER se calcula de forma análoga, pero operando a nivel de carácter en lugar de palabra:

$$\text{CER} = \frac{S_c + D_c + I_c}{N_c} \quad (2.12)$$

donde S_c , D_c e I_c representan sustituciones, eliminaciones e inserciones de caracteres, y N_c el número total de caracteres de la referencia.

La distinción entre ambas métricas resulta especialmente relevante para lenguas aglutinantes como el quechua. Debido a que una sola palabra puede incorporar múltiples sufijos, un error en un único sufijo hace que toda la palabra se contabilice como incorrecta bajo WER, penalizando de forma severa lo que a nivel de contenido es un acierto casi total. La CER, al operar sobre caracteres, refleja de manera más fiel la proporción real de la palabra correctamente reconocida, por lo que una brecha amplia entre WER y CER es indicativa de errores morfológicos superficiales más que de fallos de reconocimiento sustanciales (Joshi et al., 2020).

2.10.4. Otras métricas: COMET y METEOR

Más allá de BLEU y ChrF, existen métricas orientadas a correlacionar mejor con el juicio humano. METEOR ([Banerjee and Lavie, 2005](#)) amplía la comparación de n-gramas incorporando correspondencias por sinónimos y por raíces (*stemming*), lo que la hace menos rígida que BLEU ante variaciones léxicas. Por su parte, *Crosslingual Optimized Metric for Evaluation of Translation* (COMET) ([Rei et al., 2020](#)) adopta un enfoque distinto: en lugar de contar coincidencias superficiales, emplea representaciones neuronales preentrenadas para estimar la adecuación semántica entre la traducción, la referencia y la oración de origen, alcanzando correlaciones más altas con las evaluaciones humanas en pares de idiomas de alto recurso.

No obstante, estas métricas presentan limitaciones en el contexto de este trabajo. COMET depende de modelos preentrenados multilingües en los que el quechua está escasamente representado, por lo que sus estimaciones para este idioma no resultan confiables. Por esta razón, la evaluación de esta investigación se apoya en BLEU y ChrF, y prioriza ChrF por su robustez a nivel de caracteres y su mejor comportamiento en lenguas morfológicamente ricas, característica que la literatura asocia además con una correlación más alta con el juicio humano en este tipo de lenguas ([Popović, 2015](#)).

2.10.5. Uso complementario de BLEU y ChrF

Debido a que ambas métricas evalúan diferentes aspectos de la traducción ([Papineni et al., 2002](#); [Popović, 2015](#)), su uso conjunto permite obtener una visión más completa del desempeño del sistema ([Hendy et al., 2023](#)). BLEU ofrece una referencia consolidada y ampliamente aceptada en la literatura ([Papineni et al., 2002](#)), mientras que ChrF aporta mayor sensibilidad ante variaciones morfológicas y resulta especialmente adecuada para lenguas de bajos recursos o con estructuras aglutinantes ([Popović, 2015](#); [Joshi et al., 2020](#)).

Por ello, en esta investigación se emplean ambas métricas de manera complementaria para evaluar la calidad de las traducciones generadas ([Ortega et al., 2025](#)) y reducir las limitaciones inherentes al uso de una única medida automática ([Hendy et al., 2023](#)); el alcance de esta evaluación automática y sus límites se declaran en la Sección 1.4.

Es importante señalar que, aunque ambas métricas se expresan en rangos similares, sus valores no son directamente comparables entre sí, debido a que evalúan diferentes aspectos de la calidad de traducción.

Capítulo 3

Trabajos Relacionados

3.1. Revisión de antecedentes

A continuación se revisaron investigaciones relacionadas con el tema de la traducción y el reconocimiento de voz para idiomas, las cuales pueden ayudar a ampliar el conocimiento o contribuir al desarrollo del presente trabajo.

T. Kano, S. Sakti and S. Nakamura, (2021), Transformer-Based Direct Speech-To-Speech Translation with Transcoder, IEEE Spoken Language Technology Workshop (SLT), China.

El artículo introduce un modelo basado en Transformer para la traducción directa de habla a habla (S2ST), comparándolo con enfoques tradicionales en cascada y un modelo multi-tarea de Google. Utiliza Transcoders para mejorar la transferencia de información entre ASR, MT y TTS, optimizando la alineación entre idiomas sintácticamente distintos. Se entrenó con el corpus BTEC, evaluando pares de idiomas similares (Inglés-Español, Japonés-Coreano) y distantes (Inglés-Japonés, Japonés-Inglés), y generando datos sintéticos con Google TTS. En cuanto a configuración, el Transformer tiene 3 capas en el encoder y 6 en el decoder, 8 cabezales de atención y un tamaño oculto de 256, mientras que el RNN usa LSTM con 4 capas en el encoder y 2 en el decoder. El procesamiento de datos incluyó la extracción de Mel-spectrograms con LibROSA y la segmentación en ventanas de 50 ms. Los resultados, evaluados con BLEU y METEOR, muestran que el Transformer con Transcoder supera a los modelos en cascada y multi-tarea, especialmente en idiomas distantes, al abordar mejor la alineación y transferencia de información. Este trabajo es el primero en aplicar un Transformer para S2ST y demuestra la ventaja de su estructura sobre modelos RNN tradicionales, ofreciendo mejoras significativas en la traducción de idiomas con diferencias sintácticas.

Comentario: Esta investigación es relevante, ya que resalta la importancia de utilizar una arquitectura basada en Transformer, capaz de manejar eficientemente las complejidades de la traducción de voz entre diferentes idiomas.

Gow-Smith, E., Berard, A., Boito, M. Z., & Calapodescu, I. (2023). NAVER LABS Europe's Multilingual Speech Translation Systems for the IWSLT 2023 Low-Resource Track.

El artículo de NAVER LABS Europe presenta un enfoque eficiente en parámetros para la traducción automática del habla en entornos de bajo recurso, abordando los pares Tamasheq-Francés y Quechua-Español en el IWSLT 2023. Se utilizaron modelos de representación del habla como HuBERT, wav2vec 2.0 y XLS-R, junto con modelos de traducción como mBART50NN y NLLB-200 (1.3B y 3.3B parámetros). Se entrenó en 4 GPUs V100 (80GB) con hasta 200,000 actualizaciones, congelando el modelo de MT y ajustando solo 2-3 capas del encoder, complementadas con adaptadores y convoluciones para mejorar la adaptación. Los conjuntos de datos incluyeron 17 horas de Tamasheq-Francés y solo 2 horas de Quechua-Español, además de datos multilingües de mTEDx y TED-LIUM. El modelo alcanzó un nuevo estado del arte con 23.6 BLEU en Tamasheq-Francés (+7.7 BLEU sobre el segundo mejor participante) y 17.7 BLEU en Quechua-Español, aunque se detectó contaminación en los datos de prueba, lo que llevó a un reentrenamiento sin datos duplicados. También se probó en la tarea multilingüe de IWSLT 2021, superando propuestas previas con menos parámetros y datos. Entre sus aportes más relevantes están su arquitectura multimodal y multilingüe, su eficiencia computacional, su capacidad de traducción zero-shot (por ejemplo, hacia inglés y coreano) y su posibilidad de aprendizaje incremental, permitiendo agregar nuevos idiomas sin reentrenamiento completo.

Comentario: Este trabajo es especialmente pertinente por tres motivos concretos. Primero, emplea el mismo modelo de traducción que la presente tesis, NLLB-200 (1.3B y 3.3B), lo que respalda la elección del componente de MT. Segundo, alcanzó el estado del arte en quechua-español (17.7 BLEU) partiendo de apenas dos horas de audio paralelo, lo que evidencia que este par opera de forma inherente en un régimen de muy bajos recursos. Tercero, sus autores detectaron contaminación en el conjunto de prueba y reentrenaron sin duplicados, la misma práctica de control de fuga que se aplica en esta tesis (Sección 5.5.2).

Aimituma Suyo, F., & Churata Urtado, R. M. (2019). Conversor de voz a texto para el idioma quechua usando la herramienta de reconocimiento de voz Kaldi y una red neuronal profunda (Tesis de pregrado, Universidad Nacional de San Antonio Abad del Cusco). Facultad de Ingeniería Eléctrica, Electrónica, Informática y Mecánica, Escuela Profesional de Ingeniería Informática y de Sistemas.

La tesis evalúa modelos de ASR para el quechua Cusco-Collao mediante la herramienta Kaldi y una arquitectura DNN-HMM (Deep Neural Network - Hidden Markov Model). Se construyó un corpus de voz propio de 18 horas, balanceado por género, edad y localidad, grabando frases tomadas de diversas fuentes textuales. Se generaron recursos necesarios como diccionario fonético, colecciones textuales y fonemas. Para el modelo acústico se utilizó una red neuronal profunda con 3 capas ocultas y 512 nodos por capa, con tasa de aprendizaje de 0.002, mientras que el modelo de lenguaje se basó en 3-gramas. Aunque no se aplicaron técnicas de incremento de datos, sí se realizaron experimentos con diferentes configuraciones para ajustar los hiperparámetros. La eva-

luación se hizo mediante la métrica de precisión por palabra (WACC), obteniéndose un 59.20 % de precisión.

Comentario: Este trabajo se centra en la transcripción de audio continuo en quechua a texto en quechua. El estudio de Aimituma y Churata (2019) desarrolla un corpus y aplica un modelo DNN-HMM para el quechua Cusco-Collao. Su enfoque sirve como base para nuestro proyecto, que busca integrar modelos modernos como Transformers y modelos multilingües preentrenados en un sistema orientado a contextos médicos, permitiendo comparar enfoques y mejorar el rendimiento en tareas de ASR y traducción Quechua-Español.

Zafar, M., Castaldo, A., Nayak, P., Haque, R., Gajakos, N., & Way, A. (2024). The SETU-DCU submissions to IWSLT 2024 low-resource speech-to-text translation tasks. In Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)

El artículo presenta un estudio sobre traducción de voz a texto en idiomas de bajos recursos, específicamente irlandés-inglés y maltés-inglés, utilizando modelos avanzados de ASR y MT. Para el ASR, emplearon OpenAI Whisper (versión small, fine-tuneado en maltés), mientras que para la MT usaron mBART-50 y NLLB-200-distilled-600M, ambos ajustados en datos bilingües maltés-inglés. Se trabajó con datasets de IWSLT 2024, con 7,478 ejemplos de entrenamiento para irlandés-inglés y 7,542 para maltés-inglés. Se emplearon técnicas de preprocesamiento, incluyendo la eliminación de símbolos especiales, transformación a letras minúsculas y downsampling del audio a 16 kHz. Para evaluar el rendimiento, se emplearon métricas BLEU, ChrF y COMET, destacando el sistema cascada Whisper-small + NLLB-200 con BLEU 52.60, ChrF 72.12 y COMET 0.831 en maltés-inglés, mientras que el modelo End-to-End para irlandés-inglés mostró un desempeño limitado (BLEU 0.14, ChrF 33.05). Sus hallazgos resaltan que los sistemas End-to-End pueden superar a los modelos en cascada al evitar errores acumulativos del ASR, aunque estos últimos siguen siendo competitivos en escenarios específicos. Como aporte novedoso, el estudio demuestra mejoras en ASR para maltés mediante fine-tuning y analiza la efectividad de modelos preentrenados en traducción de idiomas de bajos recursos.

Comentario: Este trabajo es relevante para la investigación, ya que explora modelos como Whisper, mBART y NLLB en contextos de escasez de datos, técnicas de preprocesamiento aplicables al sistema de recolección en quechua y la comparación entre sistemas End-to-End y en cascada.

John E. Ortega, Rodolfo Zevallos, William Chen e Idris Abdulmumin. (2025). QUESPA Submission for the IWSLT 2025 Dialectal and Low-resource Speech Translation Task. In Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025).

El artículo “QUESPA Submission for the IWSLT 2025 Dialectal and Low-resource Speech Translation Task” presenta una evaluación de sistemas de traducción de voz para el par Quechua-Español en el contexto de recursos limitados. A diferencia de

trabajos anteriores, esta edición se centró únicamente en el enfoque unconstrained, incorporando modelos preentrenados y nuevas estrategias de aumento de datos. El conjunto de datos utilizado incluyó aproximadamente 1 hora y 40 minutos de audio paralelo para traducción de voz, cerca de 48 horas de datos ASR provenientes del corpus Siminchik, además de múltiples recursos textuales y corpus adicionales como Huqariq, SMOL, GATITOS y nuevos datos de Quechua Collao (quz). También se utilizaron corpus sintéticos y técnicas de aumento de datos mediante *nlpaug*, alcanzando un total aproximado de 96 horas de entrenamiento para algunos experimentos.

La investigación evaluó tres sistemas principales. El primero consistió en un sistema en cascada basado en ConMamba para ASR y NLLB-200 para MT. El segundo empleó Whisper Large V3 combinado igualmente con NLLB-200. Finalmente, el tercer sistema utilizó SpeechT5 ajustado directamente para ST, integrando técnicas de aumento de datos y un nuevo corpus de Quechua Collao. Para la etapa de traducción automática se realizaron experimentos adicionales con diferentes configuraciones de entrenamiento utilizando datos unconstrained, datos sintéticos y recursos externos, alcanzando hasta 167,052 oraciones paralelas después de procesos de limpieza y deduplicación.

Los resultados mostraron que el sistema basado en SpeechT5 obtuvo el mejor desempeño, alcanzando 26.7 BLEU y 48.6 ChrF, superando ampliamente los resultados obtenidos en ediciones anteriores del IWSLT. Los sistemas en cascada lograron resultados menores, obteniendo 15.0 BLEU con Whisper + NLLB y 14.8 BLEU con Mamba + NLLB. Además, los experimentos evidenciaron que el incremento de datos mejora el rendimiento únicamente cuando la calidad de los corpus adicionales es adecuada, ya que la incorporación de datos sintéticos de baja calidad redujo las métricas BLEU y ChrF en algunos escenarios.

Comentario: Este trabajo se relaciona con la presente investigación debido a que aborda la tarea de traducción de audio en quechua hacia texto en español en un entorno de bajos recursos, empleando distintos modelos de reconocimiento y traducción de voz, así como diversos datasets y estrategias de aumento de datos para mejorar el desempeño de los sistemas.

Moreno Veliz, O. (2021). The REPUcs Spanish–Quechua Submission to the Americas NLP 2021 Shared Task on Open Machine Translation. Americas NLP 2021, Perú.

Este trabajo presentó un sistema de traducción automática neuronal español–quechua desarrollado por REPUcs para la tarea compartida AmericasNLP 2021, centrado en la traducción automática para lenguas de bajos recursos. El trabajo utilizó como base los corpus proporcionados por AmericasNLP, incluyendo JW300, MINE-DU y Dict_misc, y además construyó nuevos conjuntos de datos paralelos mediante la extracción y alineación manual de poemas, canciones, vocabularios, páginas web educativas y manuales en quechua ayacuchano. Asimismo, incorporó el corpus Europarl español–inglés para realizar preentrenamiento mediante aprendizaje por transferencia. El sistema implementó modelos Transformer utilizando Fairseq, junto con segmentación subword mediante SentencePiece y BPE, entrenando inicialmente un modelo español–

inglés y posteriormente ajustándolo para español-quechua. Los resultados mostraron mejoras respecto al modelo base, alcanzando un BLEU de 2.18 y un chrF de 0.336 en la validación, además de obtener el primer lugar en el Track 2 y el tercer lugar en el Track 1 de la competencia AmericasNLP 2021, evidenciando que la incorporación de nuevos corpus y el aprendizaje por transferencia mejoraron el desempeño de la traducción automática para quechua.

Comentario: El trabajo de investigación propone la creación de un sistema de reconocimiento de voz para la traducción automática de idiomas. Este estudio resalta no solo el potencial de los métodos de aprendizaje por transferencia en la optimización de modelos de traducción para lenguas con pocos recursos, como el quechua, sino que también ilustra los desafíos inherentes al manejo de lenguas aglutinantes y la necesidad de recopilar nuevos corpus de datos. La incorporación de técnicas como la segmentación de subpalabras mediante BPE y el uso de un modelo basado en Transformers, son herramientas que pueden aplicarse al proyecto de reconocimiento de voz con el propósito de optimizar la precisión de las traducciones automáticas en tiempo real.

Zevallos, R., & Camacho, L. (2018), Siminchik: Un corpus de discursos para la preservación del quechua sureño, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), República Checa y Perú

Este artículo de investigación, presenta el desarrollo del primer corpus de habla del Quechua sureño, compuesto por 97 horas de conversaciones espontáneas, diseñado para entrenar sistemas de reconocimiento de voz. El trabajo resalta la relevancia de las tecnologías de la información para preservar lenguas en peligro de extinción como el Quechua, que carece de suficientes recursos electrónicos y digitales. Para el reconocimiento de voz, se emplearon modelos acústicos basados en Modelos de Markov Ocultos utilizando el HTK Toolkit, además de modelos de lenguaje de 5-gramas interpolados Kneser-Ney. Los autores también aplicaron técnicas de normalización de las transcripciones y mencionaron planes futuros para implementar redes neuronales en el reconocimiento de voz. A pesar de los desafíos técnicos, los experimentos iniciales arrojaron una tasa de error del 17.2%. El estudio destaca, además, la necesidad de fortalecer los derechos lingüísticos y la inversión estatal para la revitalización y preservación de las lenguas indígenas mediante su integración en el mundo tecnológico moderno.

Comentario: Los idiomas están desapareciendo a un ritmo alarmante, los derechos lingüísticos de los hablantes de la mayoría de los 7000 idiomas están en riesgo. El estudio "Siminchik", que introduce el primer corpus de habla del Quechua sureño, destaca el valor de los métodos de reconocimiento del habla para proteger los idiomas en riesgo de extinción. Este antecedente ofrece importantes aprendizajes para la investigación, que busca desarrollar un aplicativo de reconocimiento de voz aplicado a la traducción de idioma Quechua a Español.

3.2. Síntesis y posicionamiento del trabajo

Los antecedentes revisados se agrupan en tres líneas. La primera aborda el reconocimiento de voz del quechua: [Aimituma Suyo and Churata Urtado \(2019\)](#) para la variante Cusco-Collao con Kaldi, y [Cardenas et al. \(2018\)](#) con el corpus Siminchik, que constituye la base acústica de este trabajo. La segunda corresponde a la traducción de voz quechua-español en escenarios de bajos recursos: [Gow-Smith et al. \(2023\)](#) y [Ortega et al. \(2025\)](#), además de [Zafar et al. \(2024\)](#), que emplea la misma arquitectura en cascada Whisper + NLLB en otros pares de idiomas. La tercera aporta la traducción automática de texto para el quechua ([Moreno Veliz, 2021](#)) y el fundamento del uso del Transformer para la traducción de voz ([Kano et al., 2021](#)).

Respecto a estos trabajos, la presente investigación se distingue en dos aspectos. Primero, aborda el dominio médico para la traducción de voz quechua-español, no cubierto por los antecedentes anteriores, orientados a dominios generales o de habla espontánea. Segundo, realiza una comparación controlada entre el enfoque en cascada y el directo sobre un mismo conjunto de prueba, con un control explícito de fuga de datos entre entrenamiento y prueba (Sección 5.5.2), aspecto rara vez cuantificado en la literatura del par.

Finalmente, conviene subrayar que las métricas BLEU reportadas por estos trabajos no son directamente comparables entre sí ni con las de esta tesis, dado que se calculan sobre conjuntos de prueba, dominios y variantes dialectales distintos; el alcance de esta restricción se detalla entre las limitaciones del estudio (Sección 1.4).

Capítulo 4

Metodología

4.1. Enfoque y tipo de investigación

El presente estudio se enmarca dentro de un enfoque cuantitativo de tipo experimental, dado que tiene como objetivo evaluar y comparar el desempeño de diferentes modelos de ASR y ST aplicados al idioma quechua en el dominio médico ([Hernández-Sampieri and Mendoza Torres, 2018](#)).

Se adopta un diseño experimental controlado, en el cual se entrenan múltiples arquitecturas utilizando distintos conjuntos de datos y configuraciones, para posteriormente evaluar su rendimiento mediante métricas cuantitativas. Este enfoque permite analizar de manera objetiva el impacto de los datos y de las arquitecturas en la calidad de la transcripción y traducción ([Hernández-Sampieri and Mendoza Torres, 2018](#)).

Asimismo, la investigación se sitúa en el contexto aplicado, orientado a la solución de problemas concretos mediante el uso de tecnologías de IA para lenguas de bajos recursos, lo que implica desafíos adicionales como la escasez de datos, variabilidad dialectal y ausencia de recursos lingüísticos estandarizados ([Hernández-Sampieri and Mendoza Torres, 2018](#)).

4.2. Herramientas y entorno de trabajo

4.2.1. Google Colab

Google Colab (abreviatura de *Colaboratory*) es una plataforma basada en la nube desarrollada por *Google* que permite ejecutar código en lenguaje Python directamente desde un navegador web, sin necesidad de instalar software adicional en el equipo local. Proporciona acceso a recursos computacionales como unidades de procesamiento gráfico (Unidad de Procesamiento Gráfico (*Graphics Processing Unit*) (GPU)) y uni-

dades de procesamiento tensorial (TPU), lo que facilita el entrenamiento de modelos de aprendizaje profundo ([Google Research, 2026](#)). En este trabajo, Google Colab constituyó el entorno de cómputo empleado para el entrenamiento y la evaluación de los modelos, dado que el procesamiento se ejecuta en la infraestructura de servidores de Google y no depende de los recursos físicos locales.

4.2.2. Servicios complementarios de inteligencia artificial

Además del entorno de cómputo, se emplearon servicios externos de IA como apoyo en la preparación del corpus. La Interfaz de Programación de Aplicaciones (*Application Programming Interface*) (API) de *OpenAI* permite integrar modelos generativos mediante solicitudes HTTP para tareas como generación de texto, normalización y traducción multilingüe ([OpenAI, 2026](#)); en este trabajo se recurrió a modelos de la familia GPT (GPT-4o y variantes) como apoyo en la preparación textual del corpus ([OpenAI, 2024](#); [Hendy et al., 2023](#)). Asimismo, Google Traductor (*Google Translate*), servicio de traducción automática neuronal de *Google* ([Google, 2026](#)), se utilizó como uno de los sistemas de traducción automática para la generación de traducciones del corpus, práctica reportada previamente en la construcción de datasets bilingües ([Groves and Mundt, 2018](#); [Casal et al., 2023](#)).

4.3. Descripción general de la metodología

La metodología empleada en el presente trabajo integra un conjunto de etapas secuenciales que abarcan desde el diseño de la herramienta de recolección de datos hasta el análisis comparativo de los sistemas evaluados. Considerada en su totalidad, comprende seis fases: el diseño y desarrollo de la aplicación LAQLA, la recolección de datos, el preprocesamiento, la definición de configuraciones experimentales, el entrenamiento y evaluación, y el análisis de resultados. La Figura 4.1 ilustra de forma esquemática las cinco etapas centrales del flujo (recolección de datos, preprocesamiento, configuraciones experimentales, entrenamiento y evaluación); la fase inicial de desarrollo de la aplicación LAQLA, al tratarse de un trabajo previo de ingeniería de software, no se representa en el diagrama y se describe de manera independiente en la siguiente subsección.

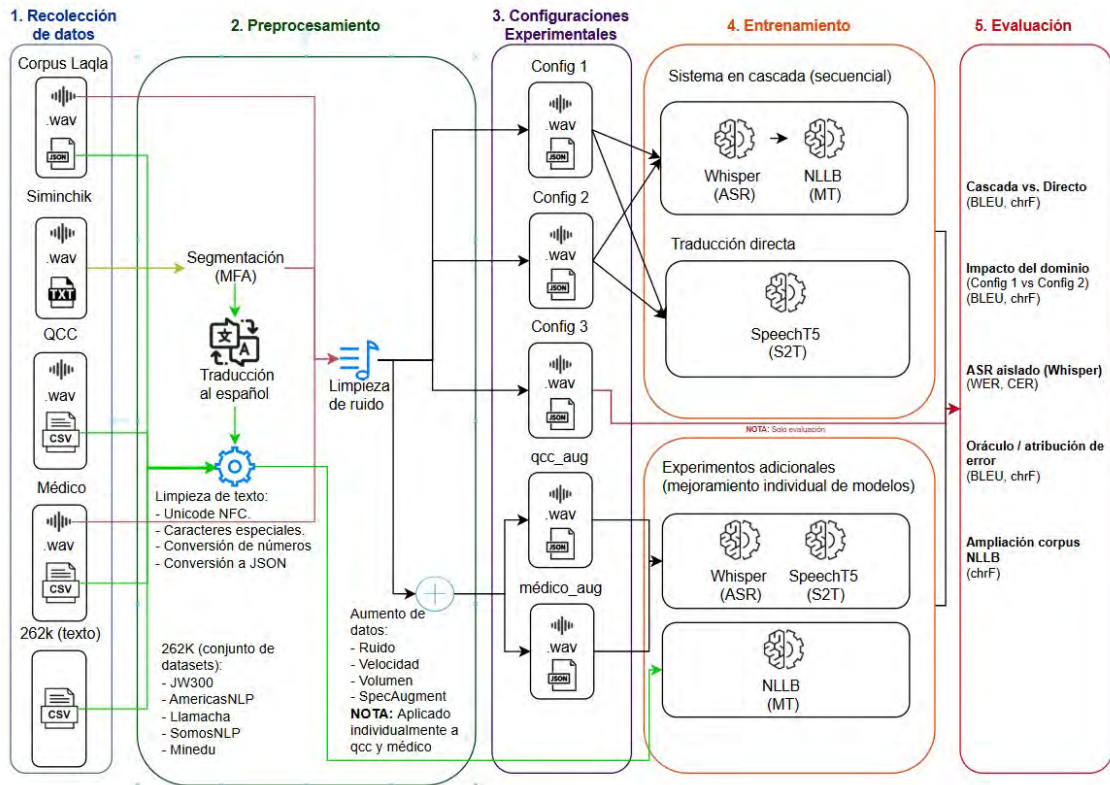


Figura 4.1: Diagrama general del flujo metodológico del trabajo de investigación.

4.3.1. Diseño de la aplicación

La primera fase del proceso consistió en el diseño y desarrollo de una aplicación móvil denominada *LAQLA*, orientada a la recolección de datos de voz en quechua en el dominio médico. Esta aplicación permite presentar textos en español a hablantes nativos de quechua, quienes graban su traducción oral al quechua. Los audios resultantes, junto con sus correspondientes textos en español, conforman el dataset *LAQLA*, un recurso de evaluación especializado en el dominio médico construido de forma propia. El desarrollo de esta aplicación constituyó un requisito previo a la etapa de recolección, ya que su arquitectura cliente-servidor garantiza la captura, validación y almacenamiento estructurado de los datos.

4.3.2. Recolección de datos

Como se aprecia en la Figura 4.1, la etapa de recolección de datos comprende la obtención de cuatro corpus diferenciados, cada uno con características propias en términos de volumen, variante dialectal y disponibilidad de recursos textuales:

- **Dataset LAQLA:** recolectado mediante la aplicación móvil desarrollada en este trabajo. Contiene grabaciones de audio en quechua y sus correspondientes traduc-

ciones de referencia en español. Los archivos se almacenan en formato de audio WAV y los textos asociados en formato JSON. Este dataset no incluye transcripciones en quechua, lo que lo convierte en un recurso utilizado exclusivamente para la evaluación final de los sistemas.

- **Dataset Siminchik:** corpus de radio en quechua sureño que contiene audio en quechua (formato WAV) junto con transcripciones textuales también en quechua, distribuidas originalmente en formato TXT. A diferencia de los demás conjuntos, no dispone de traducción al español, por lo que requiere un paso adicional de traducción automática durante el preprocesamiento.
- **Dataset qcc** (*Quechua Collao Corpus*): recurso centrado en la variante Quechua Collao que incluye audio en quechua (formato WAV), transcripciones en quechua y traducciones al español, provistas originalmente en formato CSV. Su disponibilidad de recursos alineados en ambos idiomas lo hace directamente utilizable para el entrenamiento de los modelos.
- **Dataset Médico:** corpus especializado en el dominio de la salud, no público, desarrollado por el LAAD de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC). Contiene grabaciones de audio en quechua (formato WAV) con sus respectivas transcripciones en quechua y traducciones al español, entregadas originalmente en formato CSV y construidas a partir de textos y frases representativas de situaciones relacionadas con la salud.

4.3.3. Preprocesamiento

La etapa de preprocesamiento, representada en la región central del diagrama (Figura 4.1), agrupa un conjunto de operaciones diferenciadas según las características de cada dataset.

El dataset *Siminchik* requirió un tratamiento específico en dos subetapas antes de integrarse al flujo general. En primer lugar, se llevó a cabo un proceso de segmentación que permite dividir los audios originales, cuya duración supera con frecuencia los 30 segundos, en fragmentos cortos de entre 3 y 10 segundos, más adecuados para el entrenamiento de modelos de ASR y ST. Esta segmentación se realizó mediante la herramienta Montreal Forced Aligner (MFA), que genera marcas temporales a nivel de palabra a partir del alineamiento forzado entre el audio y su transcripción. A continuación, dado que *Siminchik* no dispone de traducciones al español, se aplicó un proceso de traducción automática de las transcripciones en quechua al español utilizando el modelo GPT-5.4, seleccionado por obtener el mejor desempeño en la métrica ChrF frente a otros sistemas evaluados, siendo esta métrica la más adecuada para lenguas aglutinantes como el quechua.

Una vez aplicadas estas etapas específicas, todos los datasets (incluyendo *LAQLA*, *qcc*, *Médico* y los segmentos procesados de *Siminchik*) se sometieron a un proceso de normalización y limpieza textual común, que comprende las siguientes operaciones, tal como se indica en el diagrama:

- Normalización de los textos a formato Unicode NFC para garantizar la consistencia en la representación de caracteres.
- Eliminación de caracteres de control y espacios redundantes.
- Limpieza de caracteres especiales, incluyendo la normalización de comillas y guiones, con preservación del apóstrofe eyectivo del quechua por su relevancia fonológica.
- Conversión de los números a su forma escrita en el texto en español (mediante la librería `num2words`), con el fin de reducir la variabilidad superficial de los datos y favorecer un aprendizaje más consistente por parte de los modelos.
- Filtrado de símbolos no lingüísticos y pares de baja calidad que no cumplen con los criterios mínimos de longitud, diversidad léxica o coherencia entre idiomas.

Cabe precisar que las operaciones de limpieza descritas se implementaron mediante librerías de Python y se aplicaron de manera uniforme a todos los datasets. En lo que respecta al audio, todos los conjuntos fueron sometidos además a un proceso de normalización acústica (homogeneización de formato, recorte de silencios y ajuste de volumen) realizado igualmente con librerías de Python. No obstante, la técnica de reducción de ruido basada en el modelo MP-SENet se aplicó únicamente al dataset *LAQLA*, debido a que sus grabaciones, capturadas en condiciones no controladas a través de la aplicación móvil, presentaban un mayor nivel de ruido de fondo que el resto de corpus.

Finalmente, todos los datasets se estructuraron en formato JSON. Si bien varios corpus se distribuyen originalmente en formatos como TXT o CSV, estos presentaban numerosos errores de emparejamiento entre los audios y sus textos, ya que la correspondencia dependía de la posición del archivo o del registro dentro del CSV. Para evitar este problema, se optó por configurar los datos en formato JSON, donde cada par audio–texto se especifica de forma explícita mediante llaves, garantizando así una alineación inequívoca entre la señal acústica y su contenido textual sin depender del orden de los archivos o registros.

4.3.4. Configuraciones experimentales

La etapa de configuración experimental define los conjuntos de datos utilizados en los experimentos. Como se observa en la Figura 4.1, se establecieron tres configuraciones: dos orientadas al entrenamiento de los modelos y una tercera destinada exclusivamente a la evaluación:

- **Configuración 1:** emplea los datasets *Siminchik* y *qcc* como datos de entrenamiento, sin incorporar datos del dominio médico especializado.

- **Configuración 2:** extiende la Configuración 1 incorporando el dataset *Médico*, permitiendo evaluar el impacto de la adaptación de dominio (*domain adaptation*) en el desempeño de los modelos.
- **Configuración 3 (LAQLA):** corresponde al dataset *LAQLA* y se utilizó exclusivamente para la evaluación. Dado que este conjunto únicamente cuenta con audio en quechua y su traducción de referencia en español (pero no con transcripciones en quechua), no era posible emplearlo en el entrenamiento de los modelos. Por ello se reservó como conjunto de prueba externo, lo que permite medir el desempeño en un corpus especializado del dominio médico y en condiciones distintas a las del entrenamiento.

4.3.5. Composición dialectal y variables no controladas

Los cuatro corpus pertenecen a la rama del quechua sureño, pero no a una única variante. Los datasets qcc, Médico y LAQLA corresponden a la variante Collao, predominante en la región del Cusco, mientras que Siminchik, por tratarse de un corpus de radio del quechua sureño, combina las variantes Chanka y Collao (véase la Tabla 2.1). Dado que Siminchik se emplea como base acústica del entrenamiento y no dispone de una etiqueta de variante por enunciado, no es posible separar ni cuantificar con exactitud la proporción de habla Chanka que interviene en el entrenamiento.

En consecuencia, la variante dialectal constituye una variable no controlada del diseño experimental: el entrenamiento incluye una mezcla de Chanka y Collao (a través de Siminchik), mientras que la evaluación se realiza exclusivamente sobre corpus de variante Collao (Médico y LAQLA). Esta decisión se asume de forma consciente y responde a la escasez de recursos: prescindir de Siminchik reduciría el audio disponible para el ASR de alrededor de 42 horas a cerca de 19 horas, lo que perjudicaría el reconocimiento de voz en mayor medida que la propia heterogeneidad dialectal. Su efecto, además, se ve atenuado porque ambas variantes comparten la mayor parte de su morfología aglutinante y de su léxico dentro del quechua sureño. Se reconoce, no obstante, como una limitación del estudio (Sección 1.4), en la medida en que los resultados no garantizan el mismo desempeño en contextos monodialectales estrictos ni en otras variantes del quechua.

4.3.6. Entrenamiento y evaluación

La fase de entrenamiento y evaluación, representada en la columna derecha del diagrama (Figura 4.1), contempla dos enfoques diferenciados que compiten en la misma tarea: traducir audio en quechua a texto en español.

Sistema en cascada (Whisper + NLLB) El primer enfoque corresponde a un sistema en cascada compuesto por dos modelos entrenados de forma independiente. El mode-

lo Whisper, ajustado para la tarea de ASR, recibe el audio en quechua y genera una transcripción textual en quechua. Esta transcripción es procesada a continuación por el modelo NLLB, ajustado para la tarea de MT, que traduce el texto quechua al español. Es importante destacar que el entrenamiento de NLLB emplea exclusivamente los componentes textuales de los datasets (pares quechua-español), sin involucrar señales de audio. Ambos modelos fueron ajustados con las dos configuraciones de datos descritas, generando cuatro variantes del sistema en cascada.

Traducción directa (SpeechT5) El segundo enfoque corresponde a un modelo de traducción directa de voz. SpeechT5 recibe el audio en quechua y genera directamente el texto en español en un solo paso, sin producir una transcripción intermedia. Esta arquitectura permite evaluar si la integración de las etapas de reconocimiento y traducción en un único modelo resulta ventajosa frente al enfoque modular en el escenario de bajos recursos estudiado.

Evaluación end-to-end y evaluación adicional en LAQLA En ambos casos, la evaluación se realizó sobre el texto en español generado por el sistema completo, comparándolo con la traducción de referencia en español mediante las métricas BLEU y ChrF. Este enfoque de evaluación end-to-end, adoptado como estándar en competencias de ST como IWSLT ([Ortega et al., 2025](#)), permite medir la calidad de la traducción final sin evaluar los componentes de forma aislada.

Adicionalmente, como se indica en el diagrama, se realizó una evaluación adicional en el dataset LAQLA para ambos sistemas (cascada y traducción directa), con el objetivo de medir el desempeño en un corpus especializado en el dominio médico recolectado en condiciones diferentes a las del entrenamiento, lo que permite estimar la capacidad de generalización de los modelos.

4.3.7. Análisis de resultados

La etapa final del flujo metodológico, representada en el extremo derecho de la Figura 4.1, corresponde al análisis comparativo de los resultados obtenidos. Este análisis se estructura en torno a una tabla comparativa que consolida los valores de BLEU y ChrF para ambos sistemas (cascada y traducción directa), en sus dos configuraciones de datos (Configuración 1 y Configuración 2), evaluados sobre los dos conjuntos de prueba disponibles (dataset Médico y dataset LAQLA). A partir de esta comparación se derivan las conclusiones sobre el impacto del dataset especializado en el dominio médico, la eficiencia de cada arquitectura bajo condiciones de escasez de datos y la capacidad de generalización de los modelos desarrollados.

4.4. Recolección y construcción de datos

4.4.1. Fuentes de datos

Para el entrenamiento de los modelos se emplearon múltiples datasets provenientes tanto de iniciativas internacionales como de recolección propia. Esta combinación permite mitigar la escasez de datos y mejorar la capacidad de generalización de los modelos, estrategia ampliamente recomendada en escenarios de aprendizaje automático y lenguas de bajos recursos ([Goodfellow et al., 2016](#); [Besacier et al., 2014](#)).

4.4.1.1. Siminchik

El dataset Siminchik es uno de los principales recursos disponibles para tareas de reconocimiento automático de voz en quechua ([Cardenas et al., 2018](#)). El corpus original fue desarrollado para la preservación digital del quechua sureño y contiene aproximadamente 97.5 horas de grabaciones de audio junto con sus respectivas transcripciones textuales. El dataset incluye las variantes dialectales Quechua Chanca, hablada principalmente en Ayacucho y Apurímac, y Quechua Collao, utilizada en regiones como Cusco, Puno y parte de Bolivia ([Cardenas et al., 2018](#)).

Los audios fueron recolectados a partir de programas radiales y contienen conversaciones espontáneas, segmentos musicales, anuncios publicitarios y fragmentos de habla en español. Debido a ello, el corpus representa escenarios reales de comunicación y presenta una alta variabilidad lingüística y acústica. Los audios fueron almacenados en formato WAV monoaural con una frecuencia de muestreo de 16 kHz y precisión de 16 bits. Además, el corpus fue segmentado automáticamente en fragmentos de hasta 30 segundos para facilitar su procesamiento ([Cardenas et al., 2018](#)).

Para el presente trabajo no se utilizó el corpus completo original, sino un subconjunto de aproximadamente 48 horas de audio proporcionado en el contexto de la competición IWSLT 2025 y empleado posteriormente en el trabajo QUESPA ([Ortega et al., 2025](#)). Este subconjunto contiene audio en quechua y transcripciones textuales también en quechua, por lo que fue necesario realizar un proceso adicional de traducción de las transcripciones al español para las tareas desarrolladas en esta investigación.

Debido a la naturaleza del proceso de recolección y a las características lingüísticas del idioma, el dataset presenta diversos desafíos para el entrenamiento y evaluación de modelos de ASR. Entre los principales problemas identificados se encuentran la presencia de ruido y música de fondo provenientes de transmisiones radiales, diferencias dialectales y fonéticas entre hablantes, inconsistencias ortográficas derivadas de la limitada estandarización escrita del quechua y segmentos de audio parcialmente truncados producto de la segmentación automática ([Cardenas et al., 2018](#)). Asimismo, la naturaleza altamente aglutinante del quechua genera un vocabulario muy amplio y una elevada variabilidad léxica, lo que incrementa la complejidad de las tareas de modelado

del lenguaje y reconocimiento de voz.

Pese a los inconvenientes descritos, la inclusión de Siminchik en el entrenamiento responde a tres razones. La primera es de volumen: el ajuste de un modelo de ASR como Whisper requiere una cantidad considerable de audio, y prescindir de este corpus reduciría los datos disponibles de alrededor de 42 horas a cerca de 19, una pérdida que perjudicaría el reconocimiento de voz en mayor medida que las limitaciones de calidad que el propio corpus introduce. La segunda es de pertinencia: se trata del recurso de mayor tamaño disponible para el quechua sureño y fue construido justamente para tareas de ASR (Cardenas et al., 2018). La tercera es de precedente metodológico: el mismo subconjunto fue empleado por Ortega et al. (2025) para el ajuste de su modelo Whisper, recurriendo igualmente a traducciones automáticas de las transcripciones al español, lo que respalda su uso como recurso de entrenamiento en este escenario de bajos recursos.

4.4.1.2. Quechua Collao Corpus (qcc)

El dataset qcc es un recurso reciente utilizado en trabajos de investigación como QUESPA 2025, enfocado en el procesamiento del habla para la variante Quechua Collao (Paccotacya-Yanque et al., 2022). Este conjunto de datos fue desarrollado principalmente para tareas de reconocimiento de emociones en voz y contiene aproximadamente 15 horas de grabaciones de audio en quechua, distribuidas en 12 420 segmentos de audio (Paccotacya-Yanque et al., 2022).

A diferencia de otros corpus más generales, qcc se centra específicamente en la variante dialectal Quechua Collao, ampliamente hablada en regiones del sur del Perú y zonas aledañas. Esta especialización resulta particularmente relevante, ya que las diferencias dialectales del quechua pueden afectar significativamente el desempeño de los modelos de ASR cuando no son adecuadamente representadas en los datos de entrenamiento (Paccotacya-Yanque et al., 2022).

El dataset fue recolectado en Perú mediante sesiones de grabación controladas utilizando hablantes nativos e instructores de quechua. El corpus fue construido a partir de un guion compuesto por 378 palabras y 1692 oraciones, organizadas en nueve categorías emocionales: *happy*, *sad*, *bored*, *fear*, *sleepy*, *calm*, *excited*, *angry* y *neutral*. Asimismo, cada audio fue anotado manualmente utilizando una escala de 1 a 5 en las dimensiones emocionales *Valence*, *Arousal* y *Dominance* (Paccotacya-Yanque et al., 2022).

El corpus está compuesto principalmente por señales de audio acompañadas de sus respectivas transcripciones, traducciones y anotaciones emocionales, lo que permite su utilización en tareas de entrenamiento y evaluación de modelos de procesamiento del habla. Además, el dataset fue segmentado mediante un proceso automático basado en detección de silencios utilizando la herramienta Praat, seguido de una corrección manual para mejorar la calidad de los segmentos generados (Paccotacya-Yanque et al., 2022).

A pesar de sus aportes, el dataset presenta ciertas limitaciones técnicas. Los autores reportan que la segmentación automática no fue completamente precisa, por lo que fue necesario realizar correcciones manuales para evitar palabras truncadas, eliminar silencios innecesarios y corregir divisiones incorrectas entre segmentos de audio (Paccotacya-Yanque et al., 2022). Asimismo, al tratarse de un corpus enfocado únicamente en la variante Quechua Collao, su capacidad de generalización hacia otras variantes dialectales del quechua puede resultar limitada. Ambos aspectos se recogen entre las limitaciones del estudio (Sección 1.4).

4.4.1.3. Dataset médico

El *Dataset médico* es un recurso no público desarrollado por el LAAD de la Escuela Profesional de Ingeniería Informática y de Sistemas de la Universidad Nacional de San Antonio Abad del Cusco, en el marco de iniciativas académicas orientadas al procesamiento de lenguaje natural en contextos especializados.

Este dataset está compuesto por un total de 2474 grabaciones de audio en quechua principalmente Collao, cada una acompañada de su respectiva transcripción textual y su traducción al español. En conjunto, el corpus suma aproximadamente 4 horas de audio y fue construido a partir de grabaciones realizadas mediante lectura de frases orientadas al dominio médico.

Los audios fueron almacenados en formato WAV monoaural con una frecuencia de muestreo de 16 kHz. En cuanto a la duración de los segmentos, el 96.44 % de las grabaciones se encuentran en un rango entre 1 y 10 segundos, el 3.52 % entre 11 y 20 segundos y únicamente el 0.04 % entre 21 y 30 segundos, sin registrarse audios inválidos o con errores. Debido a que la mayoría de las grabaciones ya se encontraban dentro de los límites de duración requeridos, el dataset no necesitó procesos adicionales de segmentación.

Una característica distintiva de este corpus es su enfoque en el ámbito médico, incorporando vocabulario, expresiones y estructuras lingüísticas relacionadas con consultas y contextos de salud. Esto lo convierte en un recurso especialmente útil para el desarrollo de sistemas de ASR, MT y Voz a Texto (*Speech To Text*) (S2T) orientados a aplicaciones médicas y herramientas de apoyo para población quechua hablante.

Desde una perspectiva metodológica, este tipo de corpus especializado permite explorar escenarios de adaptación de dominio (*domain adaptation*), donde modelos entrenados con datos generales son posteriormente ajustados utilizando información específica del ámbito médico. En este sentido, el dataset puede emplearse en procesos de *fine-tuning*, favoreciendo la adaptación de los modelos hacia terminología técnica y patrones lingüísticos propios del discurso clínico (Chu et al., 2017).

A pesar de sus ventajas, el dataset presenta ciertas limitaciones. Debido a que las grabaciones fueron realizadas mediante lectura de frases previamente definidas, el corpus no captura completamente la forma de expresión natural y espontánea de los

hablantes, sino principalmente la pronunciación influenciada por el acento propio de cada participante. Esta característica, junto con el tamaño reducido del corpus, se recoge entre las limitaciones del estudio (Sección 1.4).

4.4.1.4. Dataset LAQLA (recolección propia)

Con el objetivo de generar un conjunto de datos específico para tareas de ST en el dominio médico, se desarrolló una aplicación móvil denominada *LAQLA*, orientada a la recolección de audios en quechua junto con sus respectivas traducciones en español.

La construcción de este dataset responde a la necesidad de contar con datos contextualizados en un dominio crítico como el de la salud, donde la disponibilidad de recursos en lenguas originarias es limitada. A través de la aplicación, se implementó un flujo de recolección que permite a los usuarios registrar grabaciones de voz en quechua asociadas posteriormente a su correspondiente contenido textual en español.

Desde el punto de vista metodológico, la aplicación fue diseñada para facilitar la captura de datos en condiciones reales, considerando aspectos de usabilidad y accesibilidad. El dataset recolectado está compuesto por audios almacenados en formato WAV monoaural. Asimismo, el corpus fue dividido en conjuntos de entrenamiento, validación y prueba, conformados por 526, 67 y 65 grabaciones respectivamente. Dado que LAQLA únicamente cuenta con audio en quechua y su traducción de referencia en español (pero no con transcripciones en quechua), no podía emplearse para el entrenamiento de los modelos Whisper o NLLB; en consecuencia, la totalidad del corpus quedó disponible como conjunto de evaluación externo. Debido a que la partición de prueba original resultaba reducida (65 grabaciones) para obtener una estimación estable de las métricas, la evaluación final se realizó sobre la partición de entrenamiento validada, compuesta por 433 grabaciones previamente revisadas por los expertos lingüistas. El uso de esta partición no compromete la validez del experimento, ya que LAQLA no participó en ninguna etapa del entrenamiento de los modelos, y un conjunto de mayor tamaño proporciona una estimación más robusta del desempeño.

En relación con la duración de los segmentos, la mayoría de las grabaciones presentan duraciones cortas. En el conjunto de entrenamiento, el 72.24 % de los audios se encuentran entre 2 y 5 segundos, mientras que el 12.54 % restante supera los 5 segundos. En el conjunto de validación, el 86.56 % de las grabaciones se concentra entre 2 y 5 segundos, y en el conjunto de prueba el 95.38 % también se ubica dentro de ese rango. No se registraron grabaciones inválidas o con errores en ninguno de los conjuntos.

El dataset generado mediante *LAQLA* está enfocado principalmente en el dominio médico, incorporando vocabulario especializado, expresiones clínicas y estructuras lingüísticas propias de dicho dominio, construidas a partir de textos y frases diseñadas para representar situaciones relacionadas con la salud, grabadas por hablantes nativos de la variante Quechua Collao de la región del Cusco. Esta característica lo convierte en un recurso relevante para tareas de adaptación de dominio, así como para el entrenamiento y evaluación de modelos de traducción de voz en el ámbito de la salud ([Joshi](#)

et al., 2020).

Durante el proceso de recolección, a los participantes se les mostraban textos en español correspondientes a interacciones médicas y posteriormente se les solicitaba grabar su traducción oral al quechua. Sin embargo, uno de los principales inconvenientes observados fue que, en algunos casos, los participantes no realizaban una traducción estrictamente literal del contenido presentado, sino que proporcionaban explicaciones más amplias o interpretaciones propias del mensaje original. Esto generó variabilidad semántica entre el texto fuente y la grabación producida, incrementando la complejidad del alineamiento entre ambas representaciones lingüísticas, condición que se declara entre las limitaciones del estudio (Sección 1.4).

Asimismo, debido a que las grabaciones fueron realizadas por distintos hablantes en entornos no completamente controlados, el corpus presenta variaciones en pronunciación, entonación y estilo de habla, características que permiten aproximar el estilo lingüístico propio del dominio médico.

Aplicación móvil Se desarrolló una aplicación móvil en Android Studio utilizando el lenguaje de programación Java. La aplicación presenta al usuario un texto en español, el cual debe ser leído y traducido oralmente al quechua, permitiendo así la generación de datos alineados entre ambas lenguas (Interfaz de login A.1).

La interfaz incorpora un botón de grabación que permite capturar la voz del usuario, así como la funcionalidad para reproducir el audio antes de su envío, garantizando una verificación básica por parte del participante (Ver anexo A.2: Pantalla de grabación de audios). Una vez validada la grabación, esta es enviada al servidor junto con los metadatos asociados, tales como el texto fuente, identificador del usuario y otros atributos relevantes.

Arquitectura del sistema El sistema propuesto sigue una arquitectura cliente-servidor compuesta por tres módulos principales que interactúan de manera coordinada para la recolección, almacenamiento y gestión de los datos (Ver imagen 4.2: Arquitectura de la aplicación).

La aplicación móvil (cliente Android) se comunica con un servidor backend implementado en FastAPI, el cual se encarga de gestionar las solicitudes de carga, consulta y validación de los audios. Por su parte, toda la información recolectada es almacenada en una base de datos PostgreSQL (Metadatos) y en cuanto a los archivos de audios son almacenados en MinIO, garantizando persistencia y organización estructurada de los datos (Ramírez, 2024; PostgreSQL Global Development Group, 2026; MinIO Inc., 2026; Tanenbaum and Van Steen, 2017).

En detalle, los componentes del sistema son los siguientes:

- **Aplicación móvil (cliente Android):** Responsable de la captura de audio, pre-

sentación de textos en español, envío de grabaciones al servidor, así como funcionalidades de reproducción y validación básica por parte del usuario. Además, permite el registro de información de los participantes.

- **Servidor backend (FastAPI):** Gestiona la lógica de negocio del sistema, incluyendo la recepción de audios, validación de datos, autenticación de usuarios (por ejemplo, mediante cuentas de Google) y provisión de endpoints para la interacción con el cliente.
- **Base de datos (PostgreSQL - MinIO):** Almacena de manera estructurada las grabaciones, transcripciones, traducciones, metadatos de los audios y registros de usuarios, así como información relacionada con procesos de validación.

La Figura 4.2 muestra la arquitectura general del sistema de recolección de datos, donde se ilustra la interacción entre los distintos componentes.

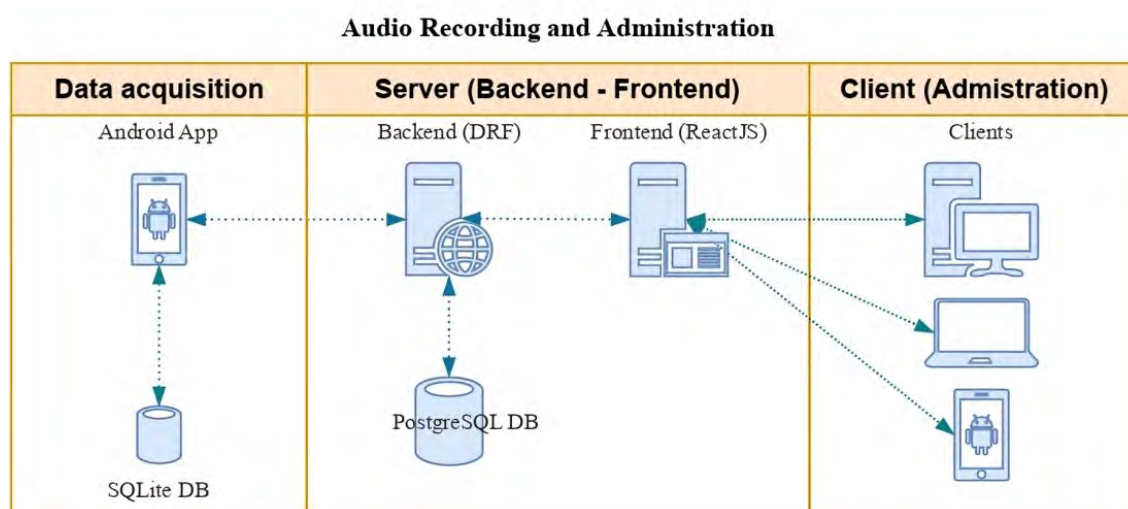


Figura 4.2: Arquitectura del sistema de recolección de audios.

Proceso de recolección La recolección de datos se llevó a cabo mediante la participación voluntaria de hablantes nativos de quechua, quienes contribuyeron con grabaciones de audio generadas a partir de textos previamente definidos en español.

Los textos utilizados fueron diseñados en el contexto de consultas médicas, con el objetivo de construir un corpus orientado a aplicaciones en el ámbito de la salud, particularmente en tareas de ASR y ST.

Cada participante realizó la lectura y traducción oral de los textos utilizando la aplicación móvil desarrollada, lo que garantizó uniformidad en el proceso de captura. Posteriormente, los audios recolectados fueron sometidos a un proceso de validación llevado a cabo por dos expertos lingüistas en quechua, quienes verificaron la correspondencia semántica entre el audio en quechua y el texto original en español (Ver

anexo A.3: Pantalla de login). Este paso resultó fundamental para asegurar la calidad lingüística del dataset y reducir la presencia de errores de traducción o ambigüedades.

Como resultado de este proceso, se obtuvo un conjunto de datos estructurado en tres subconjuntos: 526 pares para entrenamiento, 65 pares para prueba (*test*) y 67 pares para validación (*validation*). Cada registro incluye el audio y la traducción correspondiente (Ver anexo B: Estructura del dataset). Cabe precisar que ninguno de estos subconjuntos se empleó para el entrenamiento de los modelos, debido a que LAQLA no dispone de las transcripciones en quechua necesarias; la totalidad del corpus quedó reservada para la evaluación externa de los sistemas. Para dicha evaluación se utilizó la partición de entrenamiento validada (433 grabaciones tras el filtrado de duplicados), por ofrecer un conjunto de mayor tamaño y, por tanto, una estimación más estable de las métricas que la reducida partición de prueba de 65 grabaciones.

4.4.1.5. Consideraciones éticas en el tratamiento de los datos

Dado que el dataset *LAQLA* se construyó a partir de la voz de personas, la recolección se sujetó a un conjunto de resguardos éticos y de protección de datos según los puntos establecidos por la Ley N.º 29733 (consentimiento informado, finalidad, proporcionalidad y seguridad), Ley de Protección de Datos Personales del Perú. Estos resguardos se incorporaron directamente en el diseño de la aplicación móvil y de la infraestructura del servidor.

Antes de iniciar la grabación, se informó a cada participante sobre el objetivo de la investigación, el tipo de información que se registraría y el uso que se daría a las grabaciones. Se estableció de manera explícita que los audios y sus traducciones se emplearían con fines exclusivamente académicos y de investigación, sin ningún tipo de explotación comercial a terceros ajenos al proyecto. Asimismo, se comunicó que la participación era voluntaria.

La aplicación incorpora dos puntos de verificación de consentimiento en la pantalla de registro del entrevistado, que operan como modalidades alternativas y equivalentes: el registro del participante y la habilitación de la captura de audio requieren que al menos una de ellas se encuentre marcada (Ver anexo A.4: Pantalla de registro del entrevistado y consentimientos). El primero corresponde al *consentimiento por documento*, que registra la firma de un formato de consentimiento informado en formato PDF, en el cual se detallan los objetivos del estudio, el tratamiento de los datos y los derechos del participante. El segundo corresponde al *consentimiento rápido*, mediante el cual el participante expresa oralmente su aceptación y esta declaración queda registrada como una grabación de voz asociada al registro. Esta segunda modalidad se incorporó para atender las condiciones reales del trabajo de campo, donde parte de los hablantes nativos pueden presentar limitaciones de lectura en español o dificultades para completar documentación escrita.

El proceso de registro fue diseñado bajo el principio de minimización de datos: únicamente se solicitan atributos no identificativos de carácter sociodemográfico y lin-

güístico, relevantes para caracterizar el corpus, tales como el distrito, la provincia y el departamento de procedencia. No se registran nombres, apellidos, documentos de identidad, números telefónicos, direcciones ni ningún otro dato que permita la identificación directa del participante. Cada entrevistado se asocia a un identificador correlativo generado por el sistema, y los archivos de audio se almacenan bajo una nomenclatura secuencial independiente de la identidad del hablante, por lo que el dataset resultante no contiene información que permita reconstruir la identidad de las personas que participaron.

Toda la información recolectada se transmite desde la aplicación móvil hacia el servidor y se conserva en la infraestructura descrita en la Figura 4.2, donde los metadatos residen en una base de datos PostgreSQL y los archivos de audio en un almacenamiento de objetos MinIO configurado con *buckets* privados.

El acceso a los datos almacenados se realiza únicamente a través de los *endpoints* del servidor API desarrollado en FastAPI, que exige autenticación previa del usuario. La aplicación no expone rutas públicas de descarga masiva de audios, y las operaciones de consulta y validación se encuentran limitadas por rol: los participantes solo pueden acceder a sus propias grabaciones durante la sesión de captura, mientras que la revisión del corpus completo quedó reservada al equipo de investigación y a los dos expertos lingüistas encargados de la validación semántica, quienes accedieron a los audios ya anonimizados.

4.4.2. Preprocesamiento de datos

4.4.2.1. Limpieza y normalización textual

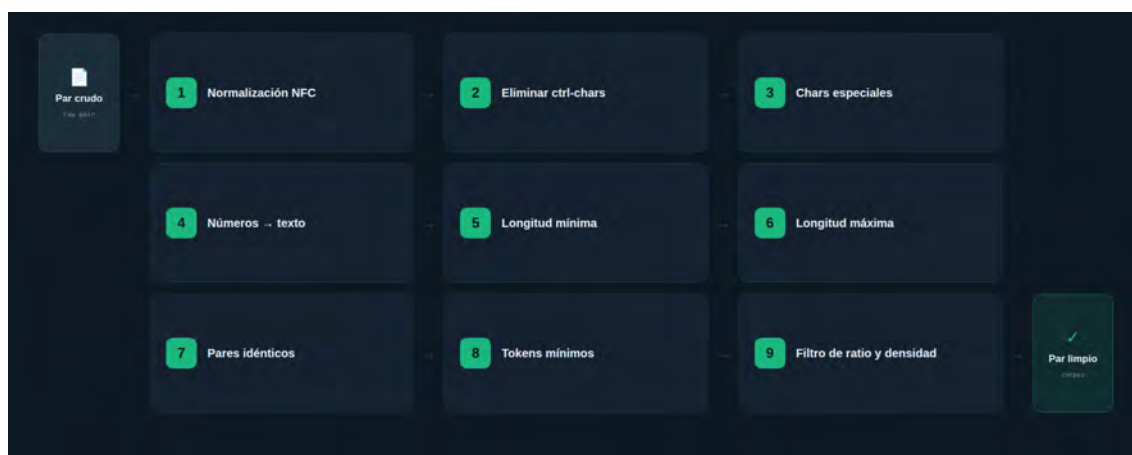


Figura 4.3: Flujo de limpieza de datos de texto.

El preprocesamiento de los datos textuales constituye una etapa crítica para garantizar la calidad del corpus y evitar que los modelos aprendan patrones erróneos o ruido presente en los datos originales (Goodfellow et al., 2016).

En una primera etapa, se realizó la normalización del texto a formato Unicode NFC, con el objetivo de asegurar la consistencia en la representación de caracteres. Asimismo, se eliminaron caracteres de control y se redujeron espacios redundantes, mejorando la uniformidad del corpus.

Posteriormente, se llevó a cabo una limpieza de caracteres especiales, que incluyó la normalización de comillas tipográficas y guiones a sus equivalentes ASCII, así como la eliminación de caracteres invisibles, tales como *zero-width spaces* y *non-breaking spaces*. Se filtraron también símbolos no lingüísticos, conservando únicamente letras, dígitos, espacios y un conjunto controlado de signos de puntuación. En el caso del quechua, se mantuvo explícitamente el uso del apóstrofe eyectivo debido a su relevancia fonológica y lingüística.

En el caso del texto en español (lengua objetivo en tareas de traducción), se aplicó una normalización semántica adicional mediante la conversión de números a su forma escrita utilizando la librería `num2words`. Esta transformación reduce la variabilidad superficial en los datos (Koehn, 2010), favoreciendo un aprendizaje más consistente por parte de los modelos.

Adicionalmente, se implementaron criterios de filtrado para eliminar pares de baja calidad o inconsistentes (Agić and Vulić, 2019). Entre estos criterios se incluyen:

- Eliminación de pares con longitud inferior o superior a umbrales predefinidos.
- Filtrado de pares idénticos entre texto fuente y destino.
- Eliminación de ejemplos con un número insuficiente de tokens.
- Restricción en la relación de longitud entre ambos idiomas, considerando la naturaleza aglutinante del quechua.
- Filtrado de textos con baja proporción de caracteres alfabéticos, lo que suele indicar contenido ruidoso o no lingüístico.

Finalmente, los pares que cumplían con todos los criterios establecidos fueron conservados manteniendo su alineación original. Este proceso permitió obtener un corpus más limpio, coherente y adecuado para el entrenamiento de modelos de aprendizaje profundo (Goodfellow et al., 2016), contribuyendo a mejorar su desempeño y capacidad de generalización.

4.4.2.2. Segmentación de audio (Siminchik)

La segmentación de audios es el proceso mediante el cual una señal sonora continua se divide en unidades menores temporalmente delimitadas, tales como palabras, locuciones o intervalos acotados por silencios. Constituye una etapa fundamental en el reconocimiento automático del habla y en la construcción de corpus, ya que permite

eliminar silencios prolongados, preparar fragmentos homogéneos para el entrenamiento y facilitar la alineación entre el audio y el texto (Rabiner and Juang, 1993; Rousso et al., 2024).

Debido a que los audios del dataset *Siminchik* (Cardenas et al., 2018) presentan duraciones variables y, en muchos casos, superiores a las recomendadas para el entrenamiento de modelos de ASR, se implementó un proceso de segmentación automática utilizando la herramienta MFA (McAuliffe et al., 2017).

Inicialmente, se realizó un análisis exploratorio de la duración de los audios, cuyos resultados se muestran en la Tabla 4.1.

Tabla 4.1: *Distribución por rango de duración de los audios*

Rango de duración (s)	Cantidad	Porcentaje
1–10	1254	16.77 %
11–20	556	7.43 %
21–30	2963	39.62 %
31–40	2704	36.15 %
41–50	2	0.03 %
Total válidos	7479	100 %
Total errores	0	—

Nota. Distribución de audios según su duración en segundos.

En sistemas modernos de reconocimiento automático de voz (ASR) y traducción directa de voz, la segmentación del audio en fragmentos de corta duración constituye una práctica fundamental para optimizar el entrenamiento de modelos secuencia a secuencia. Estudios recientes en traducción de voz indican que los modelos presentan limitaciones al procesar audios extensos debido al incremento en la complejidad computacional y a las dificultades de alineamiento temporal entre entrada y salida Tsiamas et al. (2022); Ao et al. (2022).

Asimismo, trabajos en traducción de voz extremo a extremo muestran que la segmentación adecuada permite conservar mayor calidad de traducción y reducir la degradación del rendimiento durante inferencia y entrenamiento Tsiamas et al. (2022). En este contexto, dividir grabaciones largas en segmentos más pequeños ofrece las siguientes ventajas:

- Reduce la complejidad computacional y el consumo de memoria durante el entrenamiento.
- Facilita el procesamiento por lotes (*batching*) y mejora la estabilidad del entrenamiento.
- Mejora el alineamiento entre el audio y las transcripciones o traducciones asociadas.

Para llevar a cabo la segmentación, se utilizó MFA ([McAuliffe et al., 2017](#)), una herramienta de código abierto, desarrollada sobre el toolkit Kaldi y ampliamente utilizada para la alineación forzada entre señales de audio y sus transcripciones textuales ([Wu et al., 2023](#); [Rousso et al., 2024](#)). MFA permite generar marcas temporales precisas a nivel de palabra o fonema, basándose en modelos acústicos y diccionarios de pronunciación.

El proceso completo se estructuró en las siguientes etapas:

1. **Preparación del corpus:** Se organizó el dataset en pares audio–texto, asegurando correspondencia uno a uno entre archivos `.wav` y `.txt`. Los audios fueron normalizados a formato mono con frecuencia de muestreo de 16 kHz, y las transcripciones fueron limpiadas para garantizar consistencia.

```
corpus/  
+-- audio_001.wav  
+-- audio_001.txt  
+-- audio_002.wav  
+-- audio_002.txt
```

Figura 4.4: Estructura del corpus de audio

2. **Generación del diccionario fonético:** Debido a la ausencia de un diccionario fonético para quechua, se generó automáticamente a partir del vocabulario del corpus utilizando un enfoque grafema-fonema simplificado, donde cada palabra se representa como una secuencia de caracteres.

```
def word_to_phones(word):  
    return " ".join(list(word))  
  
def generate_dictionary(vocab_file="vocab.txt",  
    output_file="dict.txt"):  
    with open(vocab_file, "r", encoding="utf-8") as f:  
        words = [line.strip() for line in f if line.strip()  
            ()]  
  
    with open(output_file, "w", encoding="utf-8") as f:  
        for word in words:  
            phones = word_to_phones(word)  
            f.write(f"{word} {phones}\n")
```

Figura 4.5: Generación de diccionario fonético

3. **Entrenamiento del modelo acústico:** Se entrenó un modelo acústico utilizando el corpus preparado, permitiendo aprender la relación entre las características del audio y las unidades fonéticas.

```
mfa train corpus/ dict.txt modelo_quechua.zip
```

4. **Alineamiento forzado:** Se generaron archivos *TextGrid* con marcas temporales precisas para cada palabra.

```
mfa align corpus/ dict.txt modelo_quechua.zip output/
```

5. **Segmentación basada en reglas:** A partir de los archivos TextGrid, se construyeron segmentos de audio a nivel de frase mediante reglas heurísticas.

Se utilizó un script en Python (empleando `praatio` y `pydub`) para generar segmentos a partir de las palabras alineadas.

Criterios de segmentación:

- Duración mínima: 3 segundos
- Duración máxima: 10 segundos
- Pausa máxima entre palabras: 1 segundo

Reglas de construcción:

1. Se ignoran silencios (ej. `sil`, `sp`).
2. Se agrupan palabras consecutivas.
3. Se detiene el segmento si se excede la duración máxima o hay pausas largas.
4. Se descartan segmentos demasiado cortos.

Los segmentos fueron extraídos recortando los audios originales según los intervalos definidos:

- Conversión de tiempos a milisegundos.
- Recorte del audio.
- Almacenamiento en archivos individuales `.wav`.

Ejemplo de salida:

```
audio_001_0000.wav  
audio_001_0001.wav
```

Finalmente, se generó un archivo `segments.txt` que mantiene la correspondencia entre audio y texto:

`segmento.wav - transcription`

Este archivo permite integrar fácilmente los datos segmentados en pipelines de entrenamiento de modelos ASR y ST, garantizando la correcta alineación entre señal acústica y contenido textual.

Inicialmente Siminchik contaba con 7479 archivos de audios con un total de 47 horas y 55 minutos, luego del proceso de segmentación se generaron 10637 archivos segmentados con un total de 22 horas y 50 minutos de audios, esto debido a que el modelo acústico entrenado no logró procesar algunos audios con mucho ruido por lo que estos audios se obviaron reduciendo la duración total del dataset original.

4.4.2.3. Traducción automática del dataset

Con el objetivo de entrenar modelos de ST capaces de traducir directamente de audio en quechua a texto en español, fue necesario generar traducciones automáticas para los segmentos obtenidos del dataset *Siminchik* (Cardenas et al., 2018). Este proceso permitió construir un corpus alineado audio–texto–traducción, indispensable para el entrenamiento supervisado de modelos de traducción de voz (Ortega et al., 2025).

Se realizó una comparación entre distintos sistemas de MT (Hendy et al., 2023), incluyendo *Google Traductor* (Google, 2026) y modelos basados en arquitecturas GPT (OpenAI, 2024) (GPT-4o, GPT-5.4 y GPT-mini). La evaluación se llevó a cabo utilizando las métricas BLEU y ChrF, ampliamente empleadas en tareas de MT (Koehn, 2010).

Tabla 4.2: Comparación de métricas BLEU y ChrF entre distintos sistemas de traducción

Modelo	BLEU	ChrF
Google Traductor	11.6292	44.5815
GPT-4o	9.4101	45.6320
GPT-5.4	7.9319	49.0700
GPT-mini	5.7676	41.5877

Nota. Comparación de resultados obtenidos mediante las métricas BLEU y ChrF entre distintos sistemas de traducción automática.

Los resultados muestran que *Google Traductor* (Google, 2026) obtiene el mejor desempeño en términos de BLEU, mientras que el modelo *GPT-5.4* (OpenAI, 2024) presenta el mejor rendimiento según la métrica ChrF. Dado que ambas métricas capturan distintos aspectos de la calidad de traducción (Hendy et al., 2023), se consideró necesario analizar su pertinencia en el contexto del idioma quechua.

En la evaluación de traducción automática para lenguas de bajos recursos ([Joshi et al., 2020](#)), la elección de la métrica es un aspecto crítico. El quechua, al ser una lengua aglutinante, presenta una alta variabilidad morfológica, donde una sola palabra puede contener múltiples sufijos que modifican su significado.

En este contexto, métricas basadas en palabras como BLEU tienden a penalizar excesivamente variaciones superficiales ([Koehn, 2010](#)), incluso cuando la traducción es semánticamente correcta. Por el contrario, la métrica ChrF opera a nivel de caracteres, permitiendo capturar similitudes parciales entre palabras y siendo más robusta frente a variaciones morfológicas y ortográficas.

La métrica ChrF presenta una mejor correlación con evaluaciones humanas en escenarios de traducción de lenguas de bajos recursos ([Hendy et al., 2023](#)). Por esta razón, en el presente trabajo se prioriza esta métrica para la selección del modelo de traducción, optando por *GPT-5.4* ([OpenAI, 2024, 2026](#)) como sistema principal.

El proceso de traducción fue implementado mediante un sistema automatizado en Python que utiliza la API de OpenAI ([OpenAI, 2026](#)), empleando el modelo *GPT-5.4* seleccionado en el apartado anterior para traducir al español los segmentos en quechua del dataset *Siminchik*. Este sistema fue diseñado para procesar grandes volúmenes de datos de manera eficiente, garantizando robustez y trazabilidad.

Conviene precisar que estas traducciones, al generarse de forma automática, constituyen una referencia aproximada y no una versión exenta de errores. Alcanzar una calidad plena requeriría la revisión manual de la totalidad del corpus por parte de un experto bilingüe, tarea inviable en términos de tiempo y costo para el volumen de datos manejado; por ello se adoptó la traducción automática como referencia de entrenamiento, asumiendo dicha limitación (Sección 1.4).

Se realizó una revisión manual del dataset médico proporcionado por LAAD, con el objetivo de verificar la correspondencia entre audios y transcripciones. Durante este proceso se identificaron inconsistencias, tales como audios que no correspondían al contenido textual asociado.

Como resultado, se aplicaron acciones de corrección que incluyeron la reasignación de textos y la eliminación de muestras inconsistentes, garantizando así la calidad del dataset final.

Para la construcción del dataset *LAQLA*, se utilizaron únicamente los audios previamente validados por los dos expertos lingüistas en quechua. Posteriormente, estos audios fueron sometidos a un proceso de preprocesamiento acústico orientado a mejorar su calidad y uniformidad ([Besacier et al., 2014](#)).

El flujo de procesamiento incluyó las siguientes etapas:

- **Normalización de formato:** Conversión a audio mono de 16 bits mediante FFmpeg.

- **Reducción de ruido:** Aplicación del modelo MP-SENet para eliminación de ruido de fondo.
- **Recorte de silencios:** Eliminación de silencios al inicio y final de cada grabación.
- **Normalización de volumen:** Ajuste del nivel de amplitud para homogenizar las grabaciones.

Este preprocesamiento permitió obtener audios más limpios y consistentes, lo cual es fundamental para mejorar el desempeño de los modelos de reconocimiento y traducción del habla ([Besacier et al., 2014](#)).

Con el objetivo de facilitar la experimentación durante el entrenamiento utilizando múltiples conjuntos de datos, cada dataset se mantuvo de forma independiente respecto a sus particiones: entrenamiento (*train*), validación (*validation*) y prueba (*test*) ([Goodfellow et al., 2016](#)).

Esta estrategia permite configurar de manera flexible el proceso de entrenamiento, seleccionando de forma específica qué partición de cada dataset será utilizada en las distintas fases (entrenamiento, validación o evaluación) ([Goodfellow et al., 2016](#)).

Para la generación de resultados y la comparación entre experimentos, se utilizó la partición de validación del dataset *Médico*. De esta manera, se garantiza una evaluación consistente del desempeño de los diferentes modelos entrenados ([Koehn, 2010](#)).

4.5. Modelos utilizados

La selección de las arquitecturas SpeechT5, Whisper y NLLB se fundamenta en los resultados reportados en el artículo QUESPA 2025 ([Ortega et al., 2025](#)), donde estas propuestas obtuvieron el mejor desempeño en tareas de traducción de voz y traducción automática multilingüe. Dicho estudio evidenció que los enfoques de traducción directa basados en SpeechT5 superan, en determinados escenarios, a los sistemas en cascada tradicionales ([Ortega et al., 2025](#)).

Por esta razón, en el presente trabajo se eligieron estas arquitecturas como punto de partida para realizar ajuste fino (*fine-tuning*) con nuevos datasets especializados en el dominio médico, donde la precisión terminológica y la robustez del sistema son aspectos críticos.

4.5.1. ¿Por qué SpeechT5?

SpeechT5 fue seleccionado porque permite un enfoque de traducción directa de voz a texto o de voz a voz dentro de una arquitectura unificada multimodal ([Ortega](#)

et al., 2025). Esto reduce la propagación de errores entre etapas intermedias, problema común en los sistemas en cascada (Besacier et al., 2014). Además, su desempeño previo en QUESPA 2025 (Ortega et al., 2025) lo posiciona como una base sólida para adaptación al dominio médico.

4.5.2. ¿Por qué Whisper + NLLB?

La combinación Whisper + NLLB representa un sistema en cascada robusto y modular (Ortega et al., 2025). Whisper destaca por su alta precisión en ASR, incluso en condiciones acústicas adversas (Radford et al., 2023), mientras que NLLB ofrece MT multilingüe de alta calidad en una gran cantidad de idiomas (NLLB Team et al., 2022). Esta combinación permite comparar un enfoque tradicional altamente competitivo frente a un modelo de traducción directa.

4.5.3. ¿Qué ventaja comparativa tienen?

SpeechT5 presenta como principal ventaja la integración de múltiples tareas en una sola arquitectura (Ao et al., 2022), lo que puede reducir errores acumulados y mejorar la latencia del sistema.

Por otro lado, Whisper + NLLB ofrece mayor interpretabilidad y flexibilidad, ya que cada componente puede optimizarse de manera independiente: reconocimiento de voz por un lado y traducción por otro (Koehn, 2010).

La comparación entre ambas estrategias permite analizar el compromiso entre rendimiento, especialización por módulos y capacidad de adaptación al dominio médico (Ortega et al., 2025).

4.5.4. Representaciones vectoriales (embeddings) en los sistemas evaluados

Los tres modelos empleados operan internamente sobre representaciones vectoriales o *embeddings*, según el concepto definido en la Sección 2.8.2 (Bhoir et al., 2017; Jurafsky and Martin, 2019). Explicitar qué magnitud se convierte en vector dentro de cada arquitectura resulta necesario para interpretar los resultados, ya que las diferencias de desempeño entre el enfoque de traducción directa y el enfoque en cascada se originan, en buena medida, en la manera en que cada uno construye y reutiliza esos espacios de representación (Joshi et al., 2020). La Tabla 4.3 resume dicha correspondencia.

De la tabla se desprende una diferencia estructural entre ambos enfoques. En el sistema en cascada, cada módulo reutiliza un espacio de representación ya consolidado

Tabla 4.3: *Representaciones vectoriales empleadas por cada modelo y su tratamiento durante el ajuste fino*

Modelo	Qué se convierte en vector	Módulo responsable	Tratamiento durante el ajuste fino
SpeechT5	Espectrograma del audio y tokens de texto, proyectados a un mismo espacio latente compartido	<i>Speech Encoder Pre-net</i> y <i>Text Encoder Prenet</i>	Se parte de los pesos preentrenados y se reajustan con los corpus de las Configuraciones 1 y 2; el vocabulario del tokenizador se mantiene sin modificar
Whisper	Espectrograma log-Mel y tokens especiales de idioma y tarea	Capas convolucionales de entrada y matriz de embeddings del decoder	Se reajustan los pesos preentrenados sobre el audio en quechua de cada configuración
NLLB	Subpalabras generadas por SentencePiece y token de idioma (quy_Latn, spa_Latn)	Matriz de embeddings de subpalabras compartida entre los idiomas del modelo	Se reajustan los pesos preentrenados sobre los pares paralelos quechua-español, sin incorporar subpalabras nuevas al vocabulario

Nota. La correspondencia entre cada arquitectura y la unidad que se convierte en vector se establece a partir de la descripción original de los modelos (Ao et al., 2022; Radford et al., 2023; NLLB Team et al., 2022).

por un preentrenamiento masivo, acústico en el caso de Whisper y textual multilingüe en el caso de NLLB, y la comunicación entre ambos módulos se realiza mediante texto en quechua, es decir, fuera del espacio vectorial (Sperber and Paulik, 2020).

En cuanto a la estrategia de entrenamiento, en ninguno de los tres casos se entrenaron representaciones vectoriales desde cero ni se amplió el vocabulario de los tokenizadores con unidades específicas del quechua. Se optó por partir de las matrices preentrenadas y permitir que el ajuste fino las reajustara (Chu et al., 2017), decisión coherente con el volumen de datos disponible: incorporar nuevas entradas de vocabulario habría exigido estimar vectores sin preentrenamiento previo a partir de un corpus reducido (Koehn and Knowles, 2017), con el consiguiente riesgo de sobreajuste (Goodfellow et al., 2016). Una consecuencia directa de esta decisión se observa en el tokenizador de SpeechT5 (Ao et al., 2022), que no contempla la letra “ñ” ni las vocales acentuadas y que, por tanto, carece de vectores asociados a esos caracteres.

4.5.5. ¿Es válida la comparación pese a la diferencia de tamaño entre modelos?

Los dos enfoques evaluados no solo difieren en su arquitectura, sino también en el número de parámetros de los modelos preentrenados que los componen. La Tabla 4.4 resume esa diferencia.

Tabla 4.4: *Número de parámetros de los modelos preentrenados empleados en cada enfoque*

Modelo (punto de control)	Rol en el sistema	Parámetros
SpeechT5 (<code>speecht5_asr</code>)	Traducción directa (ST)	≈ 152 M
Whisper (<code>whisper-small</code>)	ASR de la cascada	≈ 242 M
NLLB-200 (<code>nllb-200-1.3B</code>)	MT de la cascada	≈ 1370 M
Cascada completa (Whisper + NLLB)	Sistema en cascada	≈ 1612 M

Nota. Los valores se obtienen del recuento de pesos de los puntos de control públicos utilizados en este trabajo (Ao et al., 2022; Radford et al., 2023; NLLB Team et al., 2022). En conjunto, la cascada reúne alrededor de diez veces más parámetros que el modelo directo.

Esta asimetría es considerable y conviene explicitar por qué no invalida la comparación ni explica por sí sola los resultados obtenidos, dado que en regímenes de bajos recursos el desempeño se encuentra limitado principalmente por la disponibilidad de datos y no por la capacidad de los modelos (Joshi et al., 2020; Koehn and Knowles, 2017).

La pregunta de investigación no consiste en determinar cuál de dos redes de capacidad equivalente rinde mejor, sino cuál de dos estrategias de construcción de sistemas (cascada modular frente a traducción directa) resulta más adecuada para el par quechua-español en el dominio médico con los recursos realmente disponibles (Sperber and Paulik, 2020). Quien despliega un sistema elige entre los puntos de control preentrenados que existen para cada enfoque, y esos puntos de control tienen precisamente los tamaños de la Tabla 4.4.

Las combinaciones evaluadas son las mismas que reportan los antecedentes directos de este trabajo: QUESPA 2025 contrasta SpeechT5 ajustado para ST frente a cascadas construidas sobre NLLB-200 (Ortega et al., 2025), y el estudio de Zafar et al. (2024) enfrenta Whisper-small y NLLB-200 en esa misma disposición.

La cascada dispone de más parámetros precisamente porque descompone la tarea en dos subproblemas que cuentan con preentrenamiento masivo e independiente: Whisper se preentrena sobre cientos de miles de horas de audio multilingüe (Radford et al., 2023) y NLLB sobre corpus paralelos de doscientos idiomas (NLLB Team et al., 2022). El modelo directo, en cambio, debe aprender la correspondencia entre la señal acústica en quechua y el texto en español a partir del escaso audio paralelo existente. Esa capacidad adicional no es una ventaja concedida de forma arbitraria, sino la consecuencia natural de poder reutilizar recursos preentrenados por separado, que es justamente uno de los atributos del enfoque en cascada sometidos a evaluación (Sperber and Paulik, 2020). Suprimirla equivaldría a comparar contra una versión artificial de dicho enfoque.

Ambos sistemas se ajustaron con los mismos conjuntos de datos y las mismas particiones, y se evaluaron de extremo a extremo sobre la misma partición del corpus Médico, con idénticas métricas y con la normalización ortográfica común descrita en la

Sección 4.6.4. En consecuencia, la única variación sistemática entre ambas condiciones es aquella inherente al enfoque, tal como exige la comparación controlada entre sistemas de traducción de voz (Sperber and Paulik, 2020).

4.6. Configuración experimental

4.6.1. Diseño experimental

Con el objetivo de evaluar el desempeño de distintos enfoques para la traducción del quechua al español en el dominio médico, se definieron múltiples configuraciones experimentales que combinan diferentes modelos y conjuntos de datos (Ortega et al., 2025). Estas configuraciones permiten comparar enfoques en cascada (ASR + MT) frente a modelos de traducción directa de voz (ST) (Ortega et al., 2025).

En ambos casos, la evaluación se realizó sobre el texto en español generado por el sistema completo, comparándolo con la traducción de referencia en español mediante las métricas BLEU y ChrF. Este enfoque de evaluación end-to-end, adoptado como estándar en competencias de ST como IWSLT (Ortega et al., 2025), permite medir la calidad de la traducción final sin evaluar los componentes de forma aislada, ya que el objetivo de la tarea es la traducción completa y no la transcripción intermedia.

En particular, se evaluaron dos estrategias principales:

- **Pipeline ASR + MT:** compuesto por el modelo Whisper para ASR, seguido del modelo NLLB para traducción de texto (Ortega et al., 2025).
- **Traducción directa (SpeechT5):** modelo multimodal que realiza la traducción directamente desde audio en quechua a texto en español (Ortega et al., 2025).

4.6.2. Configuraciones evaluadas

4.6.2.1. Whisper + NLLB

En esta configuración, los modelos fueron entrenados de manera independiente:

- **Entrenamiento de NLLB (MT):**
 - Configuración 1: *Siminchik (Texto) + qcc (Texto)*
 - Configuración 2: *Siminchik (Texto) + qcc (Texto) + Médico (Texto)*
- **Entrenamiento de Whisper (ASR):**

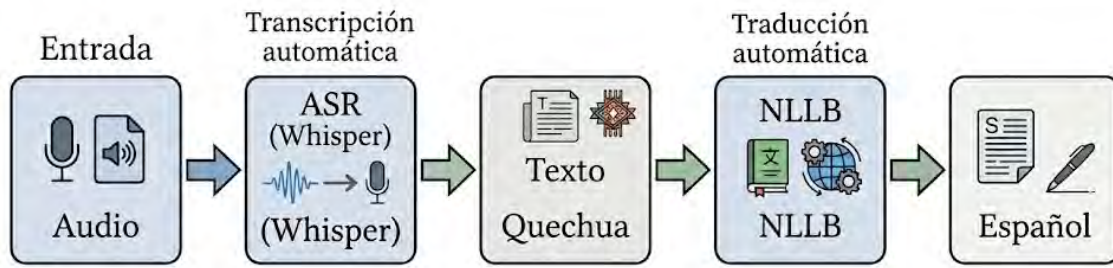


Figura 4.6: Pipeline de la arquitectura en cascada Whisper + NLLB para traducción de audios en quechua a texto

- Configuración 1: *Siminchik + qcc*
- Configuración 2: *Siminchik + qcc + Médico*

Este enfoque permite aprovechar modelos especializados para cada tarea, aunque introduce posibles errores acumulativos entre etapas.

4.6.2.2. SpeechT5



Figura 4.7: Pipeline de la arquitectura de traducción directa con el modelo SpeechT5

Se evaluó el modelo SpeechT5 en la tarea de traducción directa de voz con dos configuraciones de datos:

- **Configuración 1:**
 - *Siminchik + qcc*
- **Configuración 2:**
 - *Siminchik + qcc + Médico*

Esta estrategia busca evaluar el impacto de incorporar datos específicos del dominio médico en el desempeño del modelo.

4.6.3. Experimentos adicionales de aumento de datos y control de fuga

Además de las dos configuraciones base descritas, se llevaron a cabo experimentos adicionales orientados a mejorar el desempeño de los sistemas y a descartar posibles sesgos de la evaluación. Sus resultados completos se reportan en el Apéndice F y en el Capítulo 5; en esta sección se describe únicamente su diseño.

- **Aumento de datos acústico.** Con el fin de mitigar la escasez de audio del dominio médico, se generaron variantes aumentadas de los enunciados de entrenamiento del corpus Médico (y, en algunos experimentos, del corpus qcc) mediante técnicas de perturbación de la señal y enmascaramiento de espectrograma (SpecAugment). El aumento se aplicó exclusivamente a la partición de entrenamiento, sin alterar los conjuntos de validación y prueba, de modo que no introdujera fuga de datos. El efecto del enmascaramiento tipo SpecAugment se analiza como ablación en la Tabla F.2.
- **Ampliación del corpus paralelo para NLLB.** Para la tarea de MT se probó incorporar corpus textuales paralelos quechua-español adicionales (del orden de cientos de miles de pares), aplicando además un sobremuestreo (*oversampling*) del subconjunto médico para compensar su reducida proporción frente al resto de los datos. El objetivo era evaluar si un mayor volumen de texto mejora la traducción. Los resultados de estas variantes se reportan en la Tabla F.3.
- **Control de fuga por deduplicación.** Debido a que la partición del corpus médico no se deduplicó inicialmente y compartía algunas frases entre entrenamiento y prueba, se realizó una ablación con el corpus deduplicado para verificar si dicho solapamiento inflaba las métricas. Este control y sus resultados se detallan en la Sección 5.5.2.

4.6.4. Metodología de evaluación

La evaluación de los sistemas se realizó sobre el split de prueba del corpus Médico y sobre el corpus LAQLA, empleando el siguiente conjunto de métodos:

- **Métricas de traducción.** Se reportan BLEU (calculada con `sacrebleu`) y ChrF, esta última priorizada por su mayor correlación con el juicio humano en lenguas aglutinantes.
- **Métricas de reconocimiento.** Para el componente de ASR se calcularon la WER y la CER de Whisper frente a las transcripciones de referencia en quechua, con una normalización previa (minúsculas y sin puntuación) documentada junto a los resultados.

- **Atribución del error.** Para separar la contribución del ASR y del MT al error total, se evaluó NLLB de forma aislada alimentándolo con las transcripciones de referencia en quechua (condición oráculo), lo que establece un límite superior del sistema en cascada. El procedimiento se detalla en la Sección 5.3.2.
- **Significancia estadística.** Dada la reducida cantidad de frases de prueba, la comparación entre enfoques se acompañó de una prueba de remuestreo (*bootstrap*) pareado sobre las predicciones por oración, con el fin de determinar si las diferencias observadas son estadísticamente significativas.
- **Normalización de la evaluación.** Para garantizar una comparación justa entre el enfoque en cascada y el directo, cuyas salidas difieren en el tratamiento de las tildes y de la letra ñ, se adoptó una normalización ortográfica común, descrita en la Sección 5.1.1.

4.6.5. Consideraciones experimentales

Los experimentos se diseñaron bajo las restricciones propias de un escenario de bajos recursos –escasez de datos y variabilidad dialectal del quechua–, declaradas en la Sección 1.4 (Cardenas et al., 2018). En respuesta a ellas se combinaron múltiples datasets y se recurrió a técnicas de transferencia de aprendizaje (Goodfellow et al., 2016).

Asimismo, se buscó garantizar la comparabilidad entre configuraciones mediante el uso de conjuntos de validación y prueba consistentes en este caso se utilizó el split test del dataset Médico para la comparación de métricas, permitiendo evaluar de manera objetiva el impacto de cada configuración propuesta.

4.7. Declaración sobre el uso de herramientas de inteligencia artificial

En cumplimiento de los principios de integridad y transparencia académica, se declara el uso de herramientas basadas en IA durante la ejecución y la redacción del presente trabajo. Para ello es necesario distinguir dos planos claramente diferenciados: por un lado, la IA como objeto de estudio, que constituye el núcleo mismo de la investigación; por otro, la IA como herramienta de apoyo instrumental. En este segundo plano, las herramientas empleadas no sustituyeron en ningún caso el juicio o análisis sobre los resultados.

Se utilizó la herramienta Claude Code con el modelo Opus-5 como apoyo en la depuración y documentación de los scripts de preprocesamiento, y entrenamiento desarrollados en Python los cuales fueron revisados bajo responsabilidad.

Así mismo se utilizó la herramienta de Claude Code como apoyo en las siguientes tareas: revisión ortográfica, gramatical y de estilo; reformulación de párrafos previamente redactados para mejorar su claridad y concisión así como también para la organización de contenidos; y apoyo en la traducción del resumen al idioma inglés siendo el texto resultante revisado, corregido y validado bajo responsabilidad.

Capítulo 5

Pruebas y Resultados

5.1. Resultados experimentales

En este capítulo se presentan los resultados obtenidos por los modelos evaluados para la tarea de traducción automática de voz Quechua–Español. Se comparan dos enfoques principales: sistemas en cascada y sistemas de traducción directa.

Los experimentos fueron evaluados mediante las métricas BLEU y ChrF, ampliamente utilizadas en tareas de ST (Ortega et al., 2025; Zafar et al., 2024). En ambos enfoques cascada y traducción directa las métricas se calcularon sobre el texto en español generado por el sistema completo, comparándolo con la referencia en español. Este protocolo de evaluación end-to-end es el estándar en la evaluación de sistemas de ST para lenguas de bajos recursos, tal como lo adoptan trabajos de referencia como QUESPA (Ortega et al., 2025). BLEU mide coincidencias exactas de n-gramas respecto a una referencia, mientras que ChrF calcula la similitud a nivel de caracteres, siendo especialmente útil en lenguas morfológicamente ricas o aglutinantes como el quechua (Popović, 2015).

5.1.1. Consideraciones sobre la normalización de la evaluación

El tokenizador del modelo SpeechT5 no incluye en su vocabulario los caracteres acentuados ni la letra “ñ”, por lo que sus salidas carecen estructuralmente de estos caracteres. Para que la comparación entre el enfoque de traducción directa y el enfoque en cascada sea metodológicamente equivalente, todas las métricas reportadas en este capítulo se calcularon sobre texto sometido a una normalización ortográfica común: eliminación de tildes y sustitución de “ñ” por “n” aplicada de manera idéntica a las salidas de ambos sistemas y a las referencias. Esta decisión no favorece a ninguno de los dos enfoques, ya que las cifras resultantes varían en menos de 0,25 puntos BLEU respecto al texto sin normalizar, y elimina una fuente de sesgo ortográfico que de otro modo penalizaría al modelo directo por una limitación de su tokenizador y no por la

calidad de su traducción. La puntuación y el resto de los caracteres se conservan, dado que ambos sistemas los generan.

Asimismo, debido a que los conjuntos de prueba en escenarios de bajos recursos son reducidos (178 muestras en el dataset Médico), diferencias de pocos puntos BLEU pueden no ser distinguibles de la variabilidad del muestreo. Por ello, además de los valores puntuales, se reporta la significancia estadística de las diferencias entre enfoques mediante un procedimiento de *bootstrap* pareado con 1000 remuestras, siguiendo la práctica recomendada para la comparación de sistemas de traducción automática (Koehn, 2010).

5.2. Resultados finales

Se evaluó el sistema completo comparando los enfoques de traducción en cascada y traducción directa en el conjunto de prueba del dataset Médico.

Tabla 5.1: *Resultados obtenidos – Configuración 1*

Categoría	Sistema	BLEU	ChrF
Sistemas en cascada	Whisper + NLLB	3.00	22.10
Traducción directa	SpeechT5	0.87	16.22

Nota. Configuración 1 de datasets para entrenamiento. Para sistemas en cascada: Whisper (Siminchik + qcc) y NLLB (Siminchik (texto) + qcc (texto)). Para traducción directa: SpeechT5 (Siminchik + qcc). Métricas calculadas con la normalización descrita en la sección 5.1.1. La diferencia a favor del enfoque en cascada es estadísticamente significativa ($p < 0,01$).

Tabla 5.2: *Resultados obtenidos – Configuración 2*

Categoría	Sistema	BLEU	ChrF
Sistemas en cascada	Whisper + NLLB	5.91	24.66
Traducción directa	SpeechT5	4.68	23.67

Nota. Configuración 2 de datasets para entrenamiento. Para sistemas en cascada: Whisper (Siminchik + qcc + Médico) y NLLB (Siminchik (texto) + qcc (texto) + Médico (texto)). Para traducción directa: SpeechT5 (Siminchik + qcc + Médico). Métricas calculadas con la normalización descrita en la sección 5.1.1. En esta configuración la diferencia entre ambos enfoques no es estadísticamente significativa ($p = 0,22$).

El dataset Laqla no se incluyó en el entrenamiento con el objetivo de evaluar los modelos bajo las mismas condiciones. Esto se debe a que no cuenta con transcripciones en quechua, por lo que no puede utilizarse para entrenar los modelos Whisper o NLLB.

Sin embargo, el dataset Laqla se empleó para evaluar el desempeño del sistema completo. Los resultados se muestran en las Tablas 5.3 y 5.4.

Tabla 5.3: *Resultados de evaluación en el dataset laqla – Configuración 1*

Categoría	Sistema	BLEU	ChrF
Sistemas en cascada	Whisper + NLLB	3.71	23.38
Traducción directa	SpeechT5	0.12	15.37

Nota. Resultados obtenidos en la evaluación realizada sobre el dataset laqla utilizando la configuración 1. Métricas calculadas con la normalización descrita en la sección 5.1.1. La diferencia a favor del enfoque en cascada es estadísticamente significativa ($p < 0,001$).

Tabla 5.4: *Resultados de evaluación en el dataset laqla – Configuración 2*

Categoría	Sistema	BLEU	ChrF
Sistemas en cascada	Whisper + NLLB	4.91	24.39
Traducción directa	SpeechT5	4.85	23.95

Nota. Resultados obtenidos en la evaluación realizada sobre el dataset laqla utilizando la configuración 2. Métricas calculadas con la normalización descrita en la sección 5.1.1. En esta configuración la diferencia entre ambos enfoques no es estadísticamente significativa ($p = 0,48$).

5.3. Análisis de resultados

Los resultados evidencian que el enfoque de traducción en cascada con los modelos Whisper y NLLB en su configuración 2 alcanzó 5.91 puntos BLEU y 24.66 puntos ChrF en el split de evaluación del dataset Médico, así como 4.91 puntos BLEU y 24.39 puntos ChrF en el split de evaluación del dataset laqla. El enfoque de traducción directa con SpeechT5 obtuvo, en las mismas condiciones, 4.68 puntos BLEU y 23.67 puntos ChrF en el dataset Médico, y 4.85 puntos BLEU y 23.95 puntos ChrF en el dataset laqla.

Ahora bien, la relación entre ambos enfoques depende críticamente de la cantidad de datos de entrenamiento disponibles, y no puede resumirse en la simple superioridad de uno sobre otro. En la configuración 1, entrenada sin datos del dominio médico, el enfoque en cascada supera al directo por un margen amplio y estadísticamente significativo (3.00 frente a 0.87 puntos BLEU en el dataset Médico, $p < 0,01$; 3.71 frente a 0.12 en laqla, $p < 0,001$). Sin embargo, al incorporar el corpus médico en la configuración 2, el modelo de traducción directa reduce drásticamente esa brecha: la diferencia disminuye a poco más de un punto BLEU en el dataset Médico y a menos de un décimo de punto en laqla, y en ambos casos deja de ser estadísticamente significativa (bootstrap pareado, $p = 0,22$ y $p = 0,48$ respectivamente).

Este patrón indica que la ventaja del enfoque en cascada se concentra en el régimen de escasez extrema de datos, donde la modularidad permite que cada componente aproveche recursos preexistentes, mientras que el enfoque directo, que debe aprender de forma conjunta la correspondencia acústica y semántica, alcanza al enfoque en cascada a medida que se incorporan datos del dominio. Esta observación es coherente con lo reportado por [Ortega et al. \(2025\)](#), donde el modelo de traducción directa SpeechT5 supera a los sistemas en cascada al disponer de volúmenes de datos suficientes.

5.3.1. Impacto del dominio médico

Dado que las métricas fueron calculadas sobre el conjunto de prueba del dominio médico, se observa una mejora entre la configuración 1 y la configuración 2 en ambos enfoques. Esto indica que los modelos lograron aprender terminología y patrones lingüísticos propios del dominio clínico al incorporar datos especializados durante el entrenamiento.

La mejora cuantitativa en la evaluación del enfoque de traducción en cascada en la configuración 1 y la configuración 2 con el conjunto de datos de prueba del dataset Médico fue la siguiente:

- Incremento de 2.91 puntos BLEU.
- Incremento de 2.56 puntos ChrF.

El efecto de esta incorporación sobre el conjunto LAQLA es de menor magnitud (+1,20 BLEU y +1,01 ChrF), como recoge la Tabla 5.8.

5.3.2. Atribución del error en el enfoque en cascada

Una pregunta central para interpretar el desempeño del enfoque en cascada es determinar en qué componente se origina la mayor parte del error: en el reconocimiento de voz (ASR) o en la traducción automática (MT). Para responderla se aisló la contribución de cada etapa evaluando el módulo de traducción NLLB de forma independiente, alimentándolo con las transcripciones de referencia en quechua en lugar de las generadas por Whisper. Esta condición simula un reconocimiento de voz perfecto y establece un límite superior (*upper bound*) para el desempeño alcanzable por el enfoque en cascada. Los resultados sobre el dataset Médico en la configuración 2 se presentan en la Tabla 5.5.

Tabla 5.5: *Atribución del error en el enfoque en cascada – dataset Médico, Configuración 2*

Condición	BLEU	ChrF
NLLB con transcripción de referencia (límite superior)	8.70	25.65
Cascada completa (Whisper + NLLB)	5.91	24.66
Degradación atribuible al error de ASR	2.79	0.99

Nota. El límite superior se obtiene traduciendo con NLLB las transcripciones de referencia en quechua. La degradación por ASR es la diferencia entre ambas condiciones.

El análisis revela que, aun disponiendo de una transcripción en quechua perfecta, el módulo de traducción NLLB no supera los 8.70 puntos BLEU. Es decir, en esta

configuración el desempeño alcanzable por el enfoque en cascada está determinado principalmente por la capacidad de traducción y no por el reconocimiento de voz. La degradación introducida por los errores del ASR representa 2.79 puntos BLEU (aproximadamente un 32 % del desempeño máximo alcanzable), mientras que la brecha restante desde un sistema perfecto hasta esos 8.70 puntos corresponde íntegramente a las limitaciones del módulo de traducción.

Este hallazgo se refuerza al analizar el reconocimiento de voz de Whisper de forma aislada, que alcanza una WER de 43.80 % y una CER de 8.76 % sobre el mismo conjunto. La amplia brecha entre ambas métricas (una WER elevada frente a una CER baja) confirma que una parte sustancial de los errores del ASR son de naturaleza morfológica superficial sufijos incompletos o variaciones ortográficas sobre raíces correctamente reconocidas y no fallos de reconocimiento sustanciales. En conjunto, estos resultados sitúan en la etapa de traducción automática, y no en el reconocimiento de voz, la mayor parte del error del enfoque en cascada; la lectura de este diagnóstico y sus implicaciones para el trabajo futuro se recogen en el Capítulo 6.

5.3.3. Interpretación de BLEU y ChrF

En términos generales, los valores de BLEU fueron inferiores a los de ChrF en todos los sistemas evaluados. Esto era esperable, ya que BLEU penaliza con mayor severidad pequeñas variaciones léxicas o morfológicas.

Por otro lado, ChrF evalúa la similitud a nivel de caracteres, lo que la convierte en una métrica más adecuada para lenguas aglutinantes como el quechua, donde una misma raíz puede presentar múltiples variantes morfológicas.

Conviene situar estos valores en una escala interpretativa para evitar lecturas excesivamente optimistas. Guías de referencia difundidas en la industria, como la escala de interpretación de BLEU de Google Cloud para traducción automática ([Google Cloud, 2026](#)), consideran que un BLEU inferior a 10 corresponde a traducciones difícilmente aprovechables de forma directa, mientras que el rango de utilidad práctica suele situarse por encima de 30. Bajo ese criterio, los 5.91 puntos BLEU del mejor sistema (cascada, Configuración 2, sobre el conjunto médico) se ubican por debajo del umbral de utilidad directa: el sistema no constituye, en su estado actual, una herramienta lista para su despliegue en consultas clínicas reales.

No obstante, esa escala debe aplicarse con cautela en el presente contexto. Primero, fue calibrada sobre pares de idiomas de altos recursos y dominios generales, condiciones muy distintas a las de un par de bajos recursos, morfológicamente aglutinante y en un dominio especializado como el evaluado aquí. Segundo, BLEU se apoya en una única traducción de referencia y penaliza con severidad las variaciones morfológicas, frecuentes en quechua, por lo que tiende a subestimar la calidad real en este idioma. Por ello, el ChrF de 24.66 resulta una medida más informativa del desempeño alcanzado, al operar a nivel de caracteres y reflejar mejor las coincidencias parciales de raíces y sufijos.

La lectura de estas cifras en términos de los aportes del trabajo y de su distancia respecto al umbral de uso práctico se retoma en el Capítulo 6.

5.4. Análisis cualitativo de resultados

Además de las métricas automáticas, se realizó una inspección manual de las predicciones generadas por SpeechT5 (Ver Anexo D.1).

Se identificaron los siguientes errores frecuentes:

- Omisión de caracteres como la letra ñ.
- Pérdida de tildes y signos ortográficos.
- Sustitución de palabras por términos inexistentes.
- Cambios semánticos que alteran el significado original.
- Frases gramaticalmente incoherentes.

La omisión de la letra ñ se debe a la necesidad de normalizar el texto, reemplazando la letra “ñ” por “n” y eliminando tildes, debido a que estos caracteres no están incluidos en el vocabulario del tokenizador de SpeechT5.

Estos errores sugieren dificultades del modelo para representar adecuadamente aspectos fonéticos, sintácticos y semánticos del quechua cuando no se dispone de suficientes datos de entrenamiento.

En el caso del sistema en cascada, se observa que la transcripción de los textos realizada por el modelo Whisper presenta muy pocos errores en algunos caracteres o separaciones de palabras (Ver Anexos D.2).

No obstante, se observó que incluso en aquellos casos donde la transcripción generada por Whisper es correcta o muy similar a la referencia, la traducción al español no siempre resulta adecuada. Este comportamiento evidencia limitaciones propias del enfoque en cascada.

En primer lugar, el quechua es una lengua con morfología altamente aglutinante, en la que los sufijos incorporan información gramatical y semántica relevante. En este sentido, pequeñas variaciones en la transcripción como omisiones, segmentaciones incorrectas o ligeras alteraciones fonéticas pueden modificar la interpretación de la oración, afectando negativamente el proceso de traducción (Ver Anexo D.2).

Asimismo, el modelo de traducción presenta limitaciones asociadas a la baja disponibilidad de datos en quechua y a la variabilidad dialectal del idioma. Esto ocasiona

que, en algunos casos, el modelo genere traducciones literales o interpretaciones inadecuadas que no reflejan correctamente el significado original, especialmente en contextos específicos como el dominio médico.

Estos hallazgos indican que una transcripción precisa no garantiza una traducción de calidad, ya que el desempeño del sistema depende también de la correcta interpretación semántica y contextual en la etapa de traducción.

5.4.1. Matriz de análisis de errores

Para caracterizar de forma sistemática los errores de ambos enfoques, se analizó un conjunto representativo de enunciados del conjunto de prueba médico, contrastando la referencia en español con la salida de la cascada (Whisper + NLLB) y del modelo directo (SpeechT5). Las transcripciones completas en quechua de estos ejemplos se encuentran en los Anexos D.1 y D.2. Los errores se clasificaron según la siguiente taxonomía:

- **T1. Error de ASR propagado:** Whisper transcribe incorrectamente el audio en quechua y el error se traslada íntegro a la traducción.
- **T2. Pérdida de sufijos aglutinantes:** pequeñas alteraciones o segmentaciones de los sufijos quechua modifican la información gramatical o semántica antes de traducir.
- **T3. Omisión de término clínico (MT):** aun con una transcripción correcta, el módulo de traducción omite o sustituye un término médico relevante.
- **T4. Alucinación o degeneración fluida (directo):** SpeechT5 produce una salida gramaticalmente plausible en español pero semánticamente desconectada del audio, con frecuencia acompañada de repeticiones.

El impacto clínico del error se graduó en tres niveles: **benigno** (no altera el mensaje clínico), **matiz** (cambia un matiz sin alterar el diagnóstico) y **diagnóstico** (altera un síntoma o dato con potencial de modificar la interpretación clínica). La Tabla 5.6 reúne una selección de los casos más graves y representativos, priorizada hacia los de mayor impacto clínico, por lo que ilustra los errores más severos de cada enfoque y no constituye una distribución de frecuencia de los errores del conjunto completo; los identificadores corresponden al sufijo del archivo `quechua_00XXXX.wav`.

Tabla 5.6: *Matriz de análisis de errores en el conjunto de prueba médico (Configuración 2)*

	Audio Referencia (ES)	Cascada	Directo	Falla	Impacto
1064	¿Sabes escribir?	¿Sabes escribir?	Sabes escribir?	Ninguna	Benigno
0318	No, no tengo mucha hambre.	No tengo ganas de comer más.	No, no tengo más dolores.	D: T4	Diagnóstico
0347	Sí, mis pies se sienten entumecidos.	Sí, estoy con tu pierna perdida.	Sí, tengo pies de pie pies.	C: T3; D: T4	Diagnóstico
0420	¿Qué hace para que el dolor pase?	¿Qué haces para superar ese dolor?	¿Qué hace esa dolor para ese dolor?	D: T4	Matiz
0518	¿Desde cuándo arrastra la pierna?	A partir de ese momento, tu piel se pondrá seca.	¿Desde cuándo respire su pierna?	C: T1; D: T2	Diagnóstico
0781	A ti te está doliendo tu corazón.	¿Estás con mi corazón?	¿Tienes dolor en el corazón?	C: T1; T3	Diagnóstico
1376	¿Le duele la cabeza?	¿Tienes agua con nosotros?	Tiene dolor en cabeza?	C: T1	Diagnóstico
1395	Tengo mucha tos desde hace tres días.	Estoy más cansado que hace quince días.	Tiene mucho días de tres días.	C: T1; D: T3	Diagnóstico

Nota. **C** indica la falla del enfoque en cascada y **D** la del directo. El nivel de impacto corresponde al error más severo observado en la fila. Los códigos T1–T4 se definen en el texto.

Del análisis se desprenden tres observaciones. Primero, en los enunciados breves y frecuentes (por ejemplo, “¿Sabes escribir?”) ambos sistemas producen traducciones correctas. Segundo, los errores más graves de la cascada, con impacto de tipo diagnóstico, se originan casi siempre en la etapa de ASR (T1): al transcribir “arrastra la pierna” como “seca la piel” o “le duele la cabeza” como “agua con nosotros”, el síntoma clínico se pierde por completo, lo que concuerda con la atribución cuantitativa del error a los componentes (Sección 5.3.2). Tercero, el modelo directo falla por un mecanismo distinto, la alucinación fluida (T4), en la que la salida es gramatical pero ajena al contenido; un caso ilustrativo es la traducción de “No, no tengo mucha hambre” como “No, no tengo más dolores”, donde el modelo sustituye un término por otro clínicamente distinto (hambre por dolor).

Un matiz relevante es que, en algunos casos, el modelo directo preserva mejor el término clínico que la cascada, precisamente por no estar expuesto a los errores de transcripción del ASR: en “a ti te está doliendo tu corazón” o “¿le duele la cabeza?”, SpeechT5 conserva el síntoma (dolor de corazón, dolor de cabeza) mientras que la cascada lo pierde. Esto es coherente con la ventaja teórica del enfoque directo (Sección 2.7.4), que sin embargo no se traduce en una mejora agregada de las métricas debido a su tendencia a la alucinación cuando los datos son escasos.

Estas observaciones no contradicen la atribución cuantitativa de la Sección 5.3.2,

sino que operan en un plano distinto: los errores del ASR son menos frecuentes pero individualmente más graves, pues suprimen el síntoma clínico, mientras que las limitaciones del MT son más generalizadas y son las que fijan el techo de desempeño alcanzable.

En términos de frecuencia, la mayor parte de la brecha no atribuible al ASR (Sección 5.3.2) corresponde a errores del módulo de traducción (T2 y T3). En el modelo directo, la alucinación fluida (T4) constituye el modo de fallo predominante en los enunciados largos o poco frecuentes, mientras que su desempeño es aceptable en las frases cortas y de estructura frecuente.

5.4.1.1. Análisis detallado de casos

Para fundamentar la taxonomía anterior con evidencia directa, se presentan los mismos casos de la Tabla 5.6 en su forma original en quechua, junto con la transcripción producida por Whisper y el análisis técnico del error. Las formas en quechua y las transcripciones proceden de los Anexos D.2 y D.1.

Caso quechua_001064 (impacto benigno).

Quechua original	Yachankichu Qelqayta?
Transcripción (Whisper)	¿Yachankichu q'elqayta?
Traducción esperada	¿Sabes escribir?
Salida en cascada	¿Sabes escribir?
Salida directa (SpeechT5)	Sabes escribir?

Análisis técnico. Enunciado breve y frecuente. Whisper transcribe la raíz correctamente (solo regulariza *q'elqayta* con la oclusiva glotal) y NLLB produce la traducción exacta. El modelo directo también acierta, salvo la omisión del signo de apertura. No hay error clínico; el caso ilustra que en frases cortas y de estructura frecuente ambos enfoques son fiables.

Caso quechua_000318 (impacto diagnóstico; alucinación del enfoque directo).

Quechua original	Mana aswan mijuy munaynichu kachkan.
Transcripción (Whisper)	Mana aswan mikhuy munaynichu kachkan.
Traducción esperada	No, no tengo mucha hambre.
Salida en cascada	No tengo ganas de comer más.
Salida directa (SpeechT5)	No, no tengo más dolores.

Análisis técnico. Whisper transcribe correctamente (la variación *mijuy/mikhuy* corresponde a la misma raíz de “comer”). La cascada preserva el sentido, “no tengo ganas de comer” equivale a “no tengo hambre”, con impacto solo de matiz. El directo, en cambio, produce una salida fluida pero ajena al audio: sustituye la idea de “hambre” por

“dolores” (T4) e introduce un síntoma inexistente. El impacto diagnóstico es atribuible al enfoque directo.

Caso quechua_000347 (impacto diagnóstico; falla del MT con transcripción correcta).

Quechua original	Arí, chakiykita chinkaykuywanmi kachkani.
Transcripción (Whisper)	Arí, chakiykita chinkaykuywanmi kachkani.
Traducción esperada	Sí, mis pies se sienten entumecidos.
Salida en cascada	Sí, estoy con tu pierna perdida.
Salida directa (SpeechT5)	Sí, tengo pies de pie pies.

Análisis técnico. La transcripción de Whisper es *idéntica* a la referencia en quechua; aun así, la cascada falla en la etapa de traducción (T3): *chaki* (pie) y *chinka-* (perderse, adormecerse) se vierten como “pierna perdida”, se pierde el síntoma “entumecidos” y se altera la persona (“mis” por “tu”). Es la evidencia más nítida de que el cuello de botella reside en el MT, pues con ASR perfecto la traducción se degrada igualmente (coherente con la Sección 5.3.2). El directo degenera en una repetición sin contenido (“pies de pie pies”, T4).

Caso quechua_000420 (impacto de matiz).

Quechua original	¿Imatan ruwanky chay nanay pasanampaq?
Transcripción (Whisper)	¿Imatan ruwanki chay nanay pasanampaq?
Traducción esperada	¿Qué hace para que el dolor pase?
Salida en cascada	¿Qué haces para superar ese dolor?
Salida directa (SpeechT5)	¿Qué hace esa dolor para ese dolor?

Análisis técnico. Whisper transcribe casi perfecto (*ruwanky/ruwanki* es una variación ortográfica del sufijo de segunda persona *-nki*). La cascada conserva el sentido y solo cambia el matiz (“hace” por “haces”, “pase” por “superar”), sin alterar el mensaje clínico. El directo produce una salida repetitiva y agramatical (T4), aunque mantiene el tema “dolor”. Impacto de matiz.

Caso quechua_000518 (impacto diagnóstico; error de ASR propagado).

Quechua original	¿Hayk'akmanta chakiyki arrastranki?
Transcripción (Whisper)	Hayk'aqmanta ch'akiyki yarrastranki.
Traducción esperada	¿Desde cuándo arrastra la pierna?
Salida en cascada	A partir de ese momento, tu piel se pondrá seca.
Salida directa (SpeechT5)	¿Desde cuándo respire su pierna?

Análisis técnico. El enunciado original mezcla código: emplea el verbo español “arrastrar” con morfología quechua (*arrastra-nki*). Whisper introduce una oclusiva glotal espuria y transcribe *chakiyki* (tu pie) como *ch'akiyki*, forma cercana a la raíz *ch'aki*

(seco); sobre esa transcripción corrupta, la cascada produce “tu piel se pondrá seca” y pierde por completo el síntoma neurológico “arrastra la pierna” (T1). El directo conserva “pierna” pero altera el verbo (“arrastra” por “respire”, T2). Ambos incurren en impacto diagnóstico.

Caso quechua_000781 (impacto diagnóstico; el directo preserva mejor el síntoma).

Quechua original	Qam Sonqo nanaywanchkanki
Transcripción (Whisper)	¿Qan sonqonnanaywan kanki?
Traducción esperada	A ti te está doliendo tu corazón.
Salida en cascada	¿Estás con mi corazón?
Salida directa (SpeechT5)	¿Tienes dolor en el corazón?

Análisis técnico. Whisper segmenta mal la secuencia *sonqo nanay-wan-chka-nki* (corazón, doler, progresivo, segunda persona) y omite el aspecto progresivo *-chka-*, de modo que la cascada pierde el verbo “doler” y traduce “¿Estás con mi corazón?” (T1 y T3). El directo, al no depender de la transcripción, sí conserva “dolor en el corazón”. El caso ilustra la ventaja teórica del enfoque directo (Sección 2.7.4), que no se traduce en una mejora agregada por su tendencia a la alucinación.

Caso quechua_001376 (impacto diagnóstico; error de ASR propagado).

Quechua original	Umayki nanaykuchkanchu?
Transcripción (Whisper)	¿Umayki nanaykuq kanchu?
Traducción esperada	¿Le duele la cabeza?
Salida en cascada	¿Tienes agua con nosotros?
Salida directa (SpeechT5)	Tiene dolor en cabeza?

Análisis técnico. En *uma-yki nanay-ku-chka-n-chu* (tu cabeza, doler, progresivo, interrogativo) Whisper altera el sufijo progresivo *-chka-* y la raíz *nanay* (dolor); la cascada, sobre esa transcripción corrupta, produce “¿Tienes agua con nosotros?”, totalmente desconectada, y pierde tanto “cabeza” como “dolor” (T1). El directo conserva “dolor en cabeza”. Impacto diagnóstico para la cascada.

Caso quechua_001395 (impacto diagnóstico; error de ASR propagado).

Quechua original	Kimsa punchawmanta aswanmi sikikuchkani.
Transcripción (Whisper)	Kinza punchawmanta aswanmi sikikuskunani.
Traducción esperada	Tengo mucha tos desde hace tres días.
Salida en cascada	Estoy más cansado que hace quince días.
Salida directa (SpeechT5)	Tiene mucho días de tres días.

Análisis técnico. Whisper conserva “tres días” (*kimsa punchaw*, con la variante *kinza*) pero corrompe el verbo *sikikuchkani* (toser), lo que lleva a la cascada a sustituir el

síntoma “tos” por “cansado” y a errar la cifra temporal (“tres” por “quince”) (T1). El directo pierde igualmente el síntoma y degenera en “mucho días de tres días” (T3). Ambos incurren en impacto diagnóstico.

5.5. Discusión

5.5.1. Tamaño y calidad de los datos

El tamaño del dataset influye directamente en la capacidad de generalización de los modelos neuronales. En lenguas de bajos recursos, la escasez de datos limita el aprendizaje robusto de estructuras lingüísticas complejas, tal como se señala entre las limitaciones del estudio (Sección 1.4).

Asimismo, durante el entrenamiento de SpeechT5 se identificaron segmentos de audio mal contruidos, generados automáticamente mediante scripts de segmentación. Algunos de estos segmentos cortaban oraciones de forma incompleta, eliminando contexto semántico importante.

Ante esta limitación cabe preguntarse si ampliar el corpus mejoraría el desempeño. Esta hipótesis se evaluó de forma empírica por dos vías, cuyo diseño se describe en la Sección 4.6.3 y cuyos resultados se detallan en el Apéndice F. La primera consistió en aplicar aumento de datos acústico a los enunciados de entrenamiento (Tabla F.2); la segunda, en incorporar al módulo de traducción un corpus paralelo adicional de cientos de miles de pares quechua-español, con sobremuestreo del subconjunto médico (Tabla F.3). Ninguna de las dos intervenciones produjo una mejora agregada del desempeño: la configuración de máximo esfuerzo que las combina no superó a la Configuración 2 base.

El análisis de cobertura de dominio ofrece la explicación. El corpus paralelo adicional, de registro general, carece de la terminología clínica del conjunto de prueba: términos médicos frecuentes en las consultas evaluadas (por ejemplo, *glucosa*, *insulina*, *azúcar* o *presión*) no aparecen en dicho corpus, de modo que su incorporación aumentó el volumen de entrenamiento sin aportar cobertura del dominio objetivo. Esto sugiere que, en este problema, el factor limitante no es la cantidad de datos en términos absolutos, sino la disponibilidad de datos *del dominio clínico*: ampliar el corpus con texto general no mejora la traducción médica, mientras que la incorporación del corpus médico específico sí lo hizo (Configuración 2 frente a la Configuración 1).

Conviene precisar, además, que la cifra de referencia de 5.91 BLEU corresponde a la Configuración 2 base (Siminchik + qcc + Médico), no a la configuración de máximo esfuerzo con más datos, que no la superó. En consecuencia, la vía de mejora más prometedora no es agregar más texto de cualquier tipo, sino ampliar específicamente el corpus de voz del dominio médico, en línea con lo señalado en las limitaciones del estudio (Sección 1.4).

5.5.2. Control de fuga de datos entre entrenamiento y prueba

Durante la auditoría de los corpus se identificó que el dataset médico se particionó sin un proceso de deduplicación previo, lo que produjo un solapamiento parcial entre los conjuntos de entrenamiento y de prueba: algunas oraciones en quechua del conjunto de prueba aparecían también en el de entrenamiento del módulo de traducción. Tras el filtrado aplicado a NLLB, este solapamiento alcanzaba a lo sumo un 4 % del conjunto de prueba. Cabe precisar que las grabaciones de audio no se repiten entre particiones, por lo que el fenómeno afecta únicamente a la coincidencia del texto objetivo y no a la señal acústica.

Para determinar si esta fuga infló artificialmente las métricas reportadas, se reentrenó el módulo NLLB aplicando deduplicación de pares y un control explícito de fuga (eliminando del entrenamiento toda oración fuente que, tras normalización, coincidiera con una del conjunto de prueba) y se reevaluó el sistema en cascada de extremo a extremo sobre el mismo conjunto de prueba ($n = 178$). Este experimento controlado se realizó sobre la configuración de máximo esfuerzo, que incorpora un corpus paralelo adicional y aumento de datos acústicos, y que es por tanto la más expuesta a la memorización. Los resultados se presentan en la Tabla 5.7.

Tabla 5.7: *Control de fuga: desempeño del sistema en cascada con y sin deduplicación del corpus – dataset Médico ($n = 178$), configuración de máximo esfuerzo*

Condición	BLEU		ChrF	
	Con fuga	Sin fuga	Con fuga	Sin fuga
Cascada completa (Whisper + NLLB)	3.23	3.13	21.34	20.85
Límite superior (NLLB, ASR perfecto)	6.53	6.46	23.47	23.24

Nota. La columna “Con fuga” corresponde a la partición original; “Sin fuga” al módulo NLLB reentrenado con deduplicación y control de fuga. Ambas condiciones se evaluaron de forma idéntica sobre el mismo conjunto de prueba, por lo que la conclusión se sustenta en la diferencia, invariante ante la elección de normalización. Tras el reentrenamiento, la fuga verificada fue nula (0 coincidencias).

La eliminación completa de la fuga modifica el desempeño del sistema en cascada en menos de 0,10 puntos BLEU y menos de 0,50 puntos ChrF, una variación inferior a la incertidumbre del muestreo estimada mediante *bootstrap* para un conjunto de prueba de este tamaño. El límite superior del módulo de traducción se mantiene prácticamente inalterado (de 6.53 a 6.46 puntos BLEU). Esto indica que el solapamiento entre entrenamiento y prueba no fue la causa de los valores observados: el desempeño del enfoque en cascada es robusto frente al control de fuga.

5.5.3. Contrastación con estudios previos

El dominio sobre el cual se evaluaron los sistemas presentados corresponde específicamente al área médica; por ello, los resultados no son directamente comparables con trabajos como [Ortega et al. \(2025\)](#) o [Zafar et al. \(2024\)](#), ya que dichos estudios abordan el idioma quechua en dominios más generales y no en un contexto especializado. Para realizar una comparación más adecuada entre los modelos utilizados en este trabajo, se consideraron únicamente datasets que incluyeran alineación entre audio en quechua, transcripción en quechua y traducción al español.

En el trabajo de [Ortega et al. \(2025\)](#), el modelo con mejor desempeño fue SpeechT5 bajo un enfoque de traducción directa, alcanzando 26.7 puntos BLEU y 48.6 puntos ChrF. Este modelo superó a los sistemas en cascada tanto en los resultados reportados para 2025 como en los de 2024 presentados en el mismo estudio. Una de las principales diferencias en la configuración experimental del modelo SpeechT5 para el año 2025 fue la incorporación de un nuevo dataset durante el entrenamiento. Este resultado resalta la importancia de contar con mayores volúmenes de datos de buena calidad para mejorar el rendimiento de los modelos de traducción directa.

Por otro lado, en el trabajo de [Zafar et al. \(2024\)](#), los resultados experimentales mostraron que el sistema en cascada basado en Whisper + NLLB-200-600M obtuvo el mejor desempeño, alcanzando 52.6 puntos BLEU y 72.12 puntos ChrF. En contraste, el enfoque directo E2E basado en Whisper-small obtuvo únicamente 0.14 puntos BLEU y 33.05 puntos ChrF. Aunque los autores señalan que los modelos E2E presentan ventajas teóricas, como la reducción de errores intermedios y una menor latencia, estas ventajas no se reflejaron en el desempeño experimental reportado.

5.5.4. Resumen de los resultados obtenidos

Se presenta una tabla general de los resultados (Tabla 5.8). El Panel A reúne la comparación entre el enfoque en cascada y el enfoque de traducción directa en ambas configuraciones de entrenamiento y sobre los dos conjuntos de prueba; el bloque intermedio expresa el impacto de la incorporación del dominio médico como la diferencia entre configuraciones; y el Panel B recoge el diagnóstico del enfoque en cascada, es decir, el límite superior alcanzable con una transcripción perfecta y el desempeño del componente de reconocimiento de voz evaluado de forma aislada.

Tabla 5.8: *Resumen de los resultados obtenidos*

Enfoque / Sistema	Config.	Test Médico		Test Laqla	
		BLEU	ChrF	BLEU	ChrF
<i>Panel A. Sistemas de traducción de voz (evaluación end-to-end)</i>					
Cascada (Whisper + NLLB)	1	3.00	22.10	3.71	23.38
Directo (SpeechT5)	1	0.87	16.22	0.12	15.37
Cascada (Whisper + NLLB)	2	5.91	24.66	4.91	24.39
Directo (SpeechT5)	2	4.68	23.67	4.85	23.95
<i>Impacto del dominio médico (Δ = Config. 2 – Config. 1)</i>					
Δ Cascada (Whisper + NLLB)		+2,91	+2,56	+1,20	+1,01
Δ Directo (SpeechT5)		+3,81	+7,45	+4,73	+8,58
<i>Panel B. Diagnóstico del enfoque en cascada</i>					
NLLB con transcripción de referencia (límite superior)	2	8.70	25.65	—	—
Degradación atribuible al error de ASR	2	−2,79	−0,99	—	—
		WER ↓	CER ↓		
Whisper aislado (ASR)	2	43.80 %	8.76 %	—	—

Nota. La degradación atribuible al ASR corresponde a la diferencia entre la cascada completa y el límite superior (Tabla 5.5), expresada aquí con signo negativo por tratarse de una pérdida de desempeño. La WER y la CER se expresan en porcentaje y, a diferencia de BLEU y ChrF, valores menores indican mejor desempeño (↓). Los guiones indican que el análisis del Panel B se realizó únicamente sobre el conjunto de prueba médico en la Configuración 2.

Capítulo 6

Conclusiones y Trabajos Futuros

El presente trabajo evaluó y comparó dos enfoques de traducción automática de voz quechua a texto en español en el dominio médico de la región Cusco, el sistema en cascada Whisper + NLLB y el modelo de traducción directa SpeechT5. Como resultado de los experimentos, el sistema en cascada Whisper + NLLB en su Configuración 2 (entrenado con los datasets Siminchik, qcc y Médico) obtuvo el mejor desempeño en ambos conjuntos de evaluación: 5.91 puntos BLEU y 24.66 puntos ChrF en el dataset Médico, y 4.91 puntos BLEU y 24.39 puntos ChrF en el dataset LAQLA. Estos resultados responden al objetivo general del trabajo, identificando al sistema en cascada conformado por los modelos Whisper y NLLB como el enfoque más adecuado para la traducción de voz quechua-español en un escenario de bajos recursos, con la precisión de que su ventaja sobre el enfoque de traducción directa solo alcanza significancia estadística en ausencia de datos de entrenamiento del dominio médico, según se detalla más adelante.

La comparación entre configuraciones evidenció que la incorporación del dataset especializado en el dominio médico (Configuración 2 respecto a la Configuración 1) produjo una mejora de 2.91 puntos BLEU y 2.56 puntos ChrF en el sistema en cascada sobre el dataset Médico. Esto demuestra que, incluso con volúmenes de datos reducidos (aproximadamente 4 horas de grabaciones del dominio médico), la incorporación de datos especializados impacta positivamente en el desempeño del sistema. Ambos enfoques llegan, por tanto, a aprender representaciones útiles del quechua, pero requieren un ajuste fino (*fine-tuning*) sobre datos del dominio para alcanzar en él un desempeño aceptable.

El modelo de traducción directa SpeechT5 mostró una mayor sensibilidad al volumen de datos disponible. En su Configuración 1, obtuvo apenas 0.87 puntos BLEU y 16.22 puntos ChrF en el dataset Médico, y 0.12 puntos BLEU y 15.37 puntos ChrF en el dataset LAQLA. Al incorporar los datos médicos en la Configuración 2, mejoró a 4.68 y 23.67 puntos respectivamente en el dataset Médico, con lo que la brecha respecto del sistema en cascada se redujo hasta dejar de ser estadísticamente significativa (*bootstrap* pareado, $p = 0,22$ en el dataset Médico y $p = 0,48$ en LAQLA, conjunto en el que ambos enfoques obtienen 4.91 y 4.85 puntos BLEU respectivamente). La ventaja del

enfoque en cascada es, por tanto, amplia y estadísticamente significativa únicamente en la Configuración 1, es decir, en ausencia de datos de entrenamiento del dominio médico (3.00 frente a 0.87 puntos BLEU en el dataset Médico, $p < 0,01$), mientras que en la Configuración 2 ambos enfoques resultan equivalentes dentro de la incertidumbre del muestreo. Este comportamiento indica que la arquitectura modular resulta más efectiva cuando no se dispone de datos del dominio, al aprovechar el conocimiento preentrenado de cada componente de forma independiente, y que el enfoque de traducción directa requiere datos específicos del dominio para alcanzar un desempeño comparable.

El análisis de atribución del error situó el principal cuello de botella del enfoque en cascada en el módulo de traducción y no en el reconocimiento de voz: aun partiendo de una transcripción de referencia en quechua, el sistema no supera los 8.70 puntos BLEU, y solo 2.79 de los puntos perdidos son imputables al ASR, cuya WER de 43.80 % contrasta con una CER de 8.76 % propia de variaciones morfológicas superficiales. En consecuencia, los esfuerzos de mejora deben concentrarse en el fortalecimiento del componente de MT y en la ampliación de recursos paralelos quechua-español del dominio médico.

Es importante reconocer, no obstante, que los valores absolutos alcanzados se mantienen por debajo del umbral de utilidad práctica que suele considerarse para sistemas de traducción desplegados (un BLEU inferior a 10 se asocia a traducciones difícilmente aprovechables de forma directa). Por ello, la contribución de este trabajo no reside en un sistema listo para su uso clínico, sino en aportar el primer corpus de voz quechua-español del dominio médico, la primera comparación controlada entre los enfoques en cascada y directo para esta tarea, y una atribución cuantitativa que sitúa el principal cuello de botella en el módulo de traducción. Estos aportes ofrecen una base metodológica y una hoja de ruta clara para los trabajos futuros orientados a superar dicho umbral.

En relación con los objetivos específicos planteados, se concluye lo siguiente:

1. **Recolección de datos.** La aplicación LAQLA permitió recolectar 658 grabaciones en quechua Collao del dominio médico, de las cuales 433, validadas por dos expertos lingüistas y depuradas de duplicados, se emplearon en la evaluación. Al no contar con transcripciones en quechua, el corpus se reservó íntegramente como conjunto de evaluación externo y no participó en el entrenamiento de los modelos.
2. **Preprocesamiento de los datos.** La segmentación de Siminchik con MFA, su traducción automática con GPT-5.4 y una limpieza textual común permitieron integrar cuatro corpus heterogéneos en un conjunto de entrenamiento de alrededor de 42 horas. La reestructuración a formato JSON eliminó los errores de emparejamiento entre audio y texto propios de los formatos originales.
3. **Diseño de las configuraciones experimentales y del protocolo de comparación.** Se definieron tres configuraciones de datos que hicieron posible una comparación controlada entre ambos enfoques: dos de entrenamiento, idénticas salvo por la presencia del corpus Médico, lo que permitió aislar el efecto de la adaptación de

dominio como única variable manipulada; y una tercera constituida por el corpus LAQLA, reservada como conjunto de prueba externo.

6.1. Recomendaciones

Se recomienda continuar ampliando y mejorando el corpus especializado en consultas médicas, incorporando nuevos audios reales, mayor variedad dialectal y una cobertura más amplia de situaciones clínicas frecuentes. La participación de hablantes nativos y profesionales de la salud resulta clave para garantizar la calidad lingüística del recurso.

Asimismo, es recomendable fortalecer las etapas de limpieza, segmentación y normalización de datos antes del entrenamiento, debido a su impacto directo en los resultados finales. La validación manual de muestras críticas puede complementar los procesos automáticos.

En particular, se identificó que en el dataset SIMINCHIK varias traducciones en español no corresponden adecuadamente a los segmentos en quechua, lo cual se atribuye a un proceso de segmentación basado en criterios temporales simples (segmentos mayores a 3 segundos y menores a 10 segundos), sin considerar la integridad semántica de las oraciones. Se recomienda implementar un enfoque de segmentación asistido por modelos de lenguaje, donde, a partir de palabras previamente alineadas (ver Anexo E.1, Archivo TextGrid), se generen segmentos que correspondan a unidades lingüísticas completas y coherentes, manteniendo al mismo tiempo restricciones de duración adecuadas. Posteriormente, estos segmentos textuales pueden utilizarse como base para la segmentación de los audios.

Con respecto a la traducción al español del dataset Siminchik, obtenida en este trabajo mediante un modelo de propósito general (GPT-5.4), se recomienda sustituirla por una traducción elaborada por expertos en quechua. Los modelos de traducción de propósito general no se encuentran especializados en esta lengua, por lo que introducen errores léxicos y semánticos que se propagan al conjunto de entrenamiento y limitan la calidad de los modelos ajustados sobre él. Una vez disponible dicha traducción de referencia, se recomienda realizar una evaluación adicional bajo las mismas configuraciones experimentales descritas en el Capítulo 4 y contrastar las métricas obtenidas con las reportadas en el Capítulo 5, con el fin de cuantificar la mejora atribuible a la calidad de las traducciones y aislar el efecto del ruido introducido por la traducción automática.

Finalmente, para aplicaciones reales en salud, se recomienda complementar la evaluación automática con validaciones humanas realizadas por especialistas bilingües, a fin de medir la utilidad práctica, claridad y confiabilidad de las traducciones generadas.

6.2. Trabajos futuros

Como líneas de investigación futura, se plantea la ampliación del corpus utilizado, incorporando un mayor volumen de datos que incluya diversidad dialectal, variabilidad en contextos clínicos y un mayor número de hablantes, con el fin de mejorar la robustez y capacidad de generalización de los modelos desarrollados.

Asimismo, se propone la incorporación de métricas de evaluación semántica más avanzadas que permitan medir no solo la similitud superficial entre textos, sino también la preservación del significado, especialmente relevante en contextos médicos donde la precisión semántica es crítica.

Por otro lado, se sugiere profundizar en el desarrollo y aplicación de técnicas de segmentación más sofisticadas que consideren la estructura lingüística y la coherencia semántica de las oraciones, en lugar de basarse únicamente en criterios temporales. Esto permitiría mejorar la alineación entre audios y transcripciones, impactando positivamente en la calidad del entrenamiento de los modelos.

Otra línea de trabajo futuro consiste en explorar técnicas de congelamiento (*freezing*) de capas y de ajuste eficiente en parámetros, de forma análoga a la estrategia empleada por [Gow-Smith et al. \(2023\)](#). En dicho trabajo, el modelo de representación del habla se mantiene íntegramente congelado y se emplea únicamente como extractor de características, mientras que del modelo de traducción solo se ajustan las dos o tres capas inferiores del *encoder* junto con pequeños módulos adaptadores, dejando congelados el *decoder* y las representaciones vectoriales; de este modo se entrenan alrededor de 70 millones de parámetros sobre un total de 1.38 mil millones.

Finalmente, una vez aplicadas las mejoras en el entrenamiento y en los resultados de los modelos, se plantea su integración en una aplicación orientada a contextos médicos reales, que sirva como herramienta de apoyo a la comunicación entre el personal de salud y los pacientes quechuahablantes. Esta implementación debería acompañarse de una validación clínica que permita evaluar su utilidad práctica y su impacto en la calidad de la atención, contribuyendo a reducir las barreras lingüísticas en los servicios de salud de la región.

Apéndice

Apéndice A

Imágenes de la Aplicación

A.1. Interfaz de login

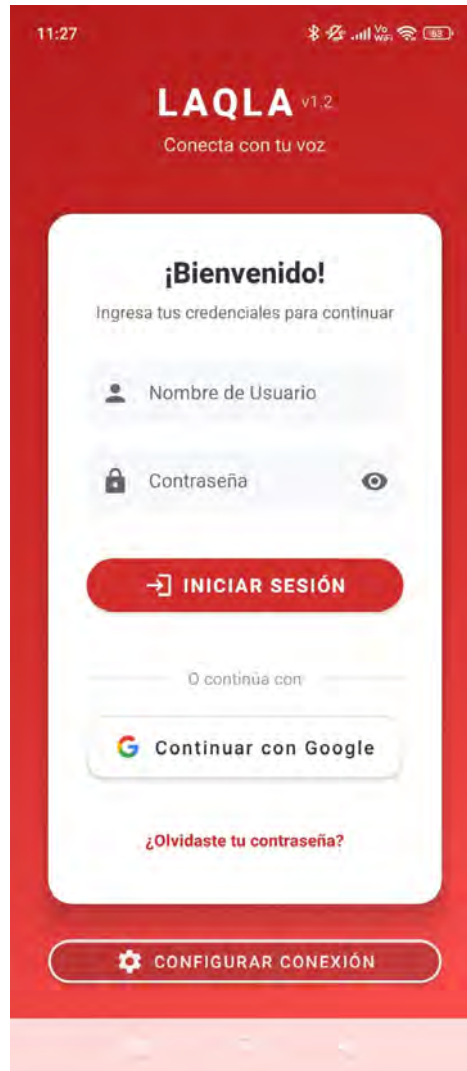


Figura A.1: Pantalla de login de la aplicación.

A.2. Pantalla de grabación de audios



Figura A.2: Interfaz de grabación de audios.

A.3. Pantalla de validación de audios



Figura A.3: Interfaz de validación de audios.

A.4. Pantalla de registro del entrevistado y consentimientos

Datos del Entrevistado
Complete la información del participante

Wanchaq

Provincia
Cusco

Departamento
Cusco

☒ **Consentimientos**

☒ Consentimiento PDF firmado

☒ Consentimiento rápido

GUARDAR ENTREVISTADO

+ CREAR NUEVO ENTREVISTADO

Entrevistado

Figura A.4: Pantalla de registro de datos no identificativos del entrevistado, con los dos puntos de verificación de consentimiento informado (consentimiento por documento firmado en PDF y consentimiento rápido registrado por voz).

Apéndice B

Estructura del Conjunto de Datos

B.1. Organización de Carpetas

```
laqla/
|-- test/
|   |-- test.json
|   \-- wav/
|       |-- quechua_000007.wav
|       \-- ...
|-- train/
|   |-- train.json
|   \-- wav/
|       |-- quechua_000000.wav
|       \-- ...
\-- valid/
    |-- valid.json
    \-- wav/
        |-- quechua_000029.wav
        \-- ...
```

test.json

```
[
  {
    "audio": "wav/audio_000000.wav",
    "quechua": "",
    "espanol": "¿Controla su peso corporal?",
    "speaker": "unknown",
    "duration": 2.799954648526077
  },

```

```
{
  "audio": "wav/audio_000001.wav",
  "quechua": "",
  "espanol": "¿Siente entumecimiento en los pies?",
  "speaker": "unknown",
  "duration": 2.7599546485260773
}
]
```

Ejemplo de contenido en archivos JSON

```
[
  {
    "audio": "wav/audio_000000.wav",
    "quechua": "",
    "espanol": "¿Controla su peso corporal?",
    "speaker": "unknown",
    "duration": 2.799954648526077
  },
  {
    "audio": "wav/audio_000001.wav",
    "quechua": "",
    "espanol": "¿Siente entumecimiento en los pies?",
    "speaker": "unknown",
    "duration": 2.7599546485260773
  },
  {
    "audio": "wav/audio_000002.wav",
    "quechua": "",
    "espanol": "¿Sangran sus encías al cepillarse?",
    "speaker": "unknown",
    "duration": 4.859954648526077
  }
]
```

Apéndice C

Repositorios del Proyecto

C.1. Aplicación Android Recolector de Audios en Quechua

Repositorio privado de la aplicación Android desarrollada para la recolección de audios en quechua:

- https://github.com/luiyiAntony/app_android_recolector_audios_quechua

C.2. Backend Recolector de Audios en Quechua

Repositorio privado del backend que gestiona la recolección y almacenamiento de los audios:

- https://github.com/luiyiAntony/backend_recolector_audios_quechua

Apéndice D

Traducciones generadas en el conjunto de prueba del dataset Médico

D.1. Traducciones del modelo SpeechT5 entrenado en la configuración 2

AUDIO : quechua_000179.wav

REF : Si, hablaremos de eso cuando este mejor.

PRED : Si, cuando les hablaremos de eso?

AUDIO : quechua_000282.wav

REF : Se aplica insulina diariamente?

PRED : Tiene que la insulina esta reciendo insulina?

AUDIO : quechua_000318.wav

REF : No, no tengo mucha hambre.

PRED : No, no tengo mas dolores?

AUDIO : quechua_000347.wav

REF : Si, mis pies se sienten entumecidos.

PRED : Si, tengo pies de pie pies?

AUDIO : quechua_000420.wav

REF : Que hace para que el dolor pase?

PRED : Que hace esa dolor para ese dolor?

AUDIO : quechua_000457.wav

REF : No tiene deseos de comer, desde cuando?

PRED : No quieres comer ni siquiera la comida de cuantos tiene?

AUDIO : quechua_000518.wav
REF : Desde cuando arrastra la pierna?
PRED : Desde cuando respire su pierna?

AUDIO : quechua_000781.wav
REF : A ti te esta doliendo tu corazon
PRED : Tienes dolor en el corazon?

AUDIO : quechua_000788.wav
REF : Como es el dolor?
PRED : Como te duele?

AUDIO : quechua_001057.wav
REF : Toda la noche?
PRED : Toda la noche?

AUDIO : quechua_001064.wav
REF : Sabes escribir?
PRED : Sabes escribir?

AUDIO : quechua_001376.wav
REF : Le duele la cabeza?
PRED : Tiene dolor en cabeza?

AUDIO : quechua_001395.wav
REF : Tengo mucha tos desde hace tres dias.
PRED : Tiene mucho dias de tres dias.

AUDIO : quechua_001488.wav
REF : Como estas ahora?
PRED : Como estas?

D.2. Transcripciones y traducciones del sistema cascada entrenado en la configuración 2

AUDIO : quechua_000179.wav
REF_QUECHUA : Ari, allinyaspayki chaymantaraq rimarisunchis.
ASR_QUECHUA : Arí, ayllaspayki chaymantaraq rimarisunchis.
REF_ESPAÑOL : Sí, hablaremos de eso cuando esté mejor.
MT_ESPAÑOL : Sí, hablemos de eso cuando quieras.

AUDIO : quechua_000282.wav
REF_QUECHUA : ¿Sapa punchaw insulinata ch'uyanchkanki?
ASR_QUECHUA : ¿Sapa punchaw insulinata chhuyanchkanki?

REF_ESPAÑOL	: Se aplica insulina diariamente?
MT_ESPAÑOL	: ¿Se inyecta insulina todos los días?
AUDIO	: quechua_000318.wav
REF_QUECHUA	: Mana aswan mijuy munaynichu kachkan.
ASR_QUECHUA	: Mana aswan mikhuy munaynichu kachkan.
REF_ESPAÑOL	: No, no tengo mucha hambre.
MT_ESPAÑOL	: No tengo ganas de comer más.
AUDIO	: quechua_000347.wav
REF_QUECHUA	: Arí, chakiykita chinkaykuywanmi kachkani.
ASR_QUECHUA	: Arí, chakiykita chinkaykuywanmi kachkani.
REF_ESPAÑOL	: Si, mis pies se sienten entumecidos.
MT_ESPAÑOL	: Sí, estoy con tu pierna perdida.
AUDIO	: quechua_000420.wav
REF_QUECHUA	: ¿Imatan ruwanky chay nanay pasanampaq?
ASR_QUECHUA	: ¿Imatan ruwanki chay nanay pasanampaq?
REF_ESPAÑOL	: Que hace para que el dolor pase?
MT_ESPAÑOL	: ¿Qué haces para superar ese dolor?
AUDIO	: quechua_000457.wav
REF_QUECHUA	: ¿Manachu mihuita munanki, hayk'aqmanta?
ASR_QUECHUA	: ¿Manachu mihuyta munanki? ¿Hayk'aqmanta?
REF_ESPAÑOL	: No tiene deseos de comer, desde cuando?
MT_ESPAÑOL	: ¿No quieres comer? ¿Cuánto tiempo?
AUDIO	: quechua_000518.wav
REF_QUECHUA	: ¿Hayk'akmanta chakiyki arrastranki?
ASR_QUECHUA	: Hayk'aqmanta ch'akiyki yarrastranki.
REF_ESPAÑOL	: Desde cuándo arrastra la pierna?
MT_ESPAÑOL	: A partir de ese momento, tu piel se pondrá seca.
AUDIO	: quechua_000781.wav
REF_QUECHUA	: Qam Sonqo nanaywanchkanki
ASR_QUECHUA	: ¿Qan sonqonanaywan kanki?
REF_ESPAÑOL	: A ti te está doliendo tu corazón
MT_ESPAÑOL	: ¿Estás con mi corazón?
AUDIO	: quechua_000788.wav
REF_QUECHUA	: ¿Imaynatatak nanasunky?
ASR_QUECHUA	: ¿Imaynatataq nanasunki?
REF_ESPAÑOL	: ¿Cómo es el dolor?
MT_ESPAÑOL	: ¿Cómo te gusta?
AUDIO	: quechua_001057.wav

REF_QUECHUA : ¿Tukuy tutachu?
ASR_QUECHUA : ¿Tukuy tutachu?
REF_ESPAÑOL : ¿Toda la noche?
MT_ESPAÑOL : ¿Toda la noche?

AUDIO : quechua_001064.wav
REF_QUECHUA : Yachankichu Qelqayta?
ASR_QUECHUA : ¿Yachankichu q'elqayta?
REF_ESPAÑOL : ¿Sabes escribir?
MT_ESPAÑOL : ¿Sabes escribir?

AUDIO : quechua_001376.wav
REF_QUECHUA : Umayki nanaykuchkanchu?
ASR_QUECHUA : ¿Umayki nanaykuq kanchu?
REF_ESPAÑOL : ¿Le duele la cabeza?
MT_ESPAÑOL : ¿Tienes agua con nosotros?

AUDIO : quechua_001395.wav
REF_QUECHUA : Kimsa punchawmanta aswanmi sikikuchkani.
ASR_QUECHUA : Kinza punchawmanta aswanmi sikikuskunani.
REF_ESPAÑOL : Tengo mucha tos desde hace tres días.
MT_ESPAÑOL : Estoy más cansado que hace quince días.

AUDIO : quechua_001488.wav
REF_QUECHUA : ¿Imaynan kunan kashianki?
ASR_QUECHUA : ¿Imaynan kunan kashanki?
REF_ESPAÑOL : ¿Cómo estás ahora?
MT_ESPAÑOL : ¿Cómo estás ahora?

Apéndice E

Alineación y segmentación

E.1. Archivo TextGrid de texto alineado

```
1 File type = "ooTextFile"
2 Object class = "TextGrid"
3
4 xmin = 0
5 xmax = 3.589
6 tiers? <exists>
7 size = 2
8 item []:
9   item [1]:
10     class = "IntervalTier"
11     name = "words"
12     xmin = 0
13     xmax = 3.589
14     intervals: size = 9
15     intervals [1]:
16       xmin = 0.0
17       xmax = 0.25
18       text = "redes"
19     intervals [2]:
20       xmin = 0.25
21       xmax = 0.48
22       text = ""
23     intervals [3]:
24       xmin = 0.48
25       xmax = 0.8
26       text = "sociales"
27     intervals [4]:
28       xmin = 0.8
29       xmax = 0.83
30       text = "y"
31     intervals [5]:
32       xmin = 0.83
33       xmax = 0.89
34       text = "el"
35     intervals [6]:
36       xmin = 0.89
37       xmax = 1.04
38       text = "canal"
39     intervals [7]:
40       xmin = 1.04
41       xmax = 1.12
42       text = "del"
```

Figura E.1: Archivo TextGrid de texto alineado.

Apéndice F

Experimentos de Ablación

Este anexo detalla los experimentos de ablación que sustentan las decisiones de configuración discutidas en el capítulo de resultados. Cada ablación aísla el efecto de un único factor (aumento de datos, dominio, tamaño del modelo o técnica de regularización) sobre la métrica final. Salvo que se indique lo contrario, las métricas del sistema directo y del componente de traducción se reportan sobre el conjunto de prueba médico ($n = 178$); las del componente ASR, sobre la validación médica durante el entrenamiento.

F.1. Sistema directo (SpeechT5): aumento de audio y dominio médico

La Tabla F.1 aísla el efecto de (i) el aumento de audio fuera de dominio (`qcc_aug`), (ii) la incorporación del dominio médico y (iii) el aumento de audio in-domain (`médico-aug`).

Tabla F.1: *Ablación del sistema directo (SpeechT5) — test médico*

Configuración de entrenamiento	BLEU	chrF
siminchik + qcc	0.87	16.22
siminchik + qcc + <code>qcc_aug</code>	0.31	13.84
siminchik + qcc + médico (<i>Config. 2</i>)	4.68	23.67
siminchik + qcc + <code>qcc_aug</code> + médico + médico-aug	8.03	26.38
siminchik + qcc + médico + médico-aug (<i>sin <code>qcc_aug</code></i>)	10.92	29.21

Hallazgos:

- **El aumento fuera de dominio perjudica al modelo directo.** Añadir `qcc_aug` sin

datos médicos reduce el chrF de 16.22 a 13.84 ($-2,4$) y el BLEU de 0.87 a 0.31. Como el decodificador no logra aterrizar en la señal acústica (*alucinación fluida*), el aumento aporta variabilidad acústica sin diversidad semántica y amplifica el comportamiento defectuoso.

- **El dominio in-domain es el factor decisivo.** Incorporar el conjunto médico eleva el chrF de 16.22 a 23.67 ($+7,5$), un salto muy superior a cualquier efecto del aumento de datos genérico.
- **El aumento in-domain (médico-aug) ayuda; el fuera de dominio (qcc_aug) sigue perjudicando.** La mejor configuración del sistema directo (siminchik + qcc + médico + médico-aug, *sin* qcc_aug) alcanza 10.92 BLEU y 29.21 chrF; añadir qcc_aug sobre esa misma receta la degrada a 8.03 y 26.38. El perjuicio del aumento fuera de dominio persiste, por tanto, incluso con el dominio médico ya presente.
- **Nota metodológica:** el modelo directo asciende con lentitud desde la zona de alucinación; la Configuración 2 requirió 39 épocas para alcanzar chrF 23.67. Por ello los valores intermedios de época temprana (chrF ≈ 12) no son indicativos del resultado final.

Nota sobre la interpretación de estas cifras. Las dos últimas configuraciones incorporan el aumento de audio in-domain (médico-aug), ausente en las tablas del capítulo de resultados; ese ingrediente explica el salto respecto a la Configuración 2. No obstante, buena parte de esa mejora sobre el test médico proviene del sobreajuste del modelo directo a las frases clínicas plantilladas, cuyo texto objetivo se solapa parcialmente entre entrenamiento y prueba (Sección 5.5.2), y no de una mayor capacidad de traducción. El enfoque en cascada no se beneficia de igual modo, pues los errores de su componente ASR rompen esa coincidencia exacta. Por ello estos valores se reportan únicamente como ablación y no modifican la recomendación del enfoque en cascada formulada en el Capítulo 6.

F.2. Componente ASR (Whisper): SpecAugment

Sin SpecAugment, `whisper-small` sobreajusta el conjunto de entrenamiento ($\sim 40k$ audios) hacia las épocas 2–3: la pérdida de entrenamiento cae hacia 0,1 mientras la de validación asciende. La activación de SpecAugment (enmascaramiento de franjas en los ejes de tiempo y frecuencia del espectrograma) reduce este sobreajuste y desplaza el mejor punto de convergencia.

Ambos valores corresponden al mejor punto de validación (época 3): sin SpecAugment ese óptimo se alcanza muy temprano y el modelo se degrada después por sobreajuste, mientras que con SpecAugment el WER desciende de 72,1 a 54,6 y el chrF sube de 69,0 a 76,1. El WER se calcula sobre el texto en minúsculas; aun así permanece elevado porque penaliza como error de palabra completa cada variación morfológica. La brecha

Tabla F.2: *Ablación de SpecAugment en el componente ASR (Whisper) — validación médica*

Configuración	WER	chrF
Sin SpecAugment	72.1	69.0
Con SpecAugment	54.6	76.1

persistente entre WER y CER ($\sim 11\%$ a nivel carácter) refleja la morfología aglutinante del quechua, no un fallo de transcripción.

F.3. Componente de traducción (NLLB): escala, curación y oversampling

La Tabla F.3 recoge tres variantes del componente de traducción sobre el dominio médico. En estos experimentos, aumentar el volumen de datos, curar el corpus o emplear un modelo 2.5 veces mayor no produjo mejoras sustanciales: los tres resultados se mantuvieron próximos entre sí, en torno a 28–29 chrF, y el oversampling in-domain del conjunto médico alcanzó el mejor valor. Los valores corresponden al chrF en validación médica durante el entrenamiento.

Tabla F.3: *Ablación del componente de traducción (NLLB) — validación médica*

Corpus	Modelo	Estrategia	chrF
262k (96 % bíblico)	1.3B	fine-tuning suave	27.9
70k curado (Cusco)	3.3B	fine-tuning suave	28.1
262k + médico	1.3B	oversampling ($\sim 4\%$)	médico 29.18

El oversampling in-domain del conjunto médico sigue la estrategia de *mixed fine-tuning* de Chu et al. (2017), en línea con el sobremuestreo de datos escasos empleado en el propio NLLB (NLLB Team et al., 2022); el factor de repetición se ajustó empíricamente.

F.3.1. Control de dominio general (FLORES-200)

Para verificar que el fine-tuning no degrada la capacidad general del modelo, se evaluó en FLORES-200 (devtest, *quy_Latn*→*spa_Latn*, $n = 1012$). Conviene precisar que FLORES-200 es un *benchmark* de dominio general (texto enciclopédico) y que su partición de quechua corresponde a la variante **Ayacucho-Chanka** (*quy_Latn*), distinta del quechua **Cusco-Collao** en el que se centra este trabajo. Por ello esta evaluación cumple únicamente la función de control de olvido catastrófico, y sus valores absolutos

no son comparables ni con los resultados del dominio médico ni con los que otros trabajos reportan sobre sus propios conjuntos de prueba (por ejemplo, los 17.7 BLEU de [Gow-Smith et al. \(2023\)](#) se miden sobre el conjunto de IWSLT, en un dominio y una distribución distintos).

Tabla F.4: *Evaluación en FLORES-200 (dominio general): base vs. fine-tuned*

Modelo	BLEU	chrF	chrF++
Base (nllb-200-1.3B)	9.05	31.76	29.92
Fine-tuned	9.52	32.87	30.92
Δ	+0.47	+1.11	+1.00

Mediante un test de *paired bootstrap* (10 000 remuestreos), la mejora en chrF (+1,11; IC 95 % [+0,85, +1,42]) es estadísticamente significativa ($p < 0,0001$). El fine-tuned supera al base también en dominio general, lo que confirma la **ausencia de olvido catastrófico**: el corpus mayoritariamente bíblico no dañó la capacidad general del modelo. En conjunto, estas ablaciones sugieren que, a esta escala y con los recursos disponibles, las estrategias exploradas no aportaron mejoras adicionales sustanciales al componente de traducción para el par quechua-español.

Bibliografía

- Instituto Nacional de Estadística e Informática (2018). Resultados definitivos de los censos nacionales 2017. 82.6 % habla castellano y 13.9 % quechua como lengua materna.
- Agić, Ž. and Vulić, I. (2019). Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of ACL*.
- Aimituma Suyo, F. and Churata Urtado, R. M. (2019). Conversor de voz a texto para el idioma quechua usando la herramienta de reconocimiento de voz kaldi y una red neuronal profunda.
- Ao, J., Wang, R., Zhou, L., Wang, C., Ren, S., Wu, Y., Liu, S., Ko, T., Li, Q., Zhang, Y., Wei, Z., Qian, Y., Li, J., and Wei, F. (2022). Specht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5723–5738.
- Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language mode. *Advances in neural information processing systems*, 13.
- Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.
- Bhoir, S., Ghorpade, T., and Mane, V. (2017). Comparative analysis of different word embedding models. *International conference on advances in computing, communication and Control*, pages 1–4.
- Cahuana Q., R. (2007). *Manual de Gramática Quechua Cusco-Collao*. Sicuani, Perú, edición revisada edition.
- Cardenas, R., Zevallos, R., Baquerizo, R., and Camacho, L. (2018). Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP 2*, page 21.
- Casal, D. et al. (2023). Using machine translation tools to build multilingual text classification datasets. *Journal of Artificial Intelligence Research*.

- Castillo, A. N. (2022). Técnicas de machine learning para procesamiento del lenguaje natural.
- Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- Google (2026). Google translate. Disponible en <https://translate.google.com>.
- Google Cloud (2026). The BLEU translation quality metric. Cloud Translation Documentation. Escala de interpretación de BLEU (0–100): < 10 casi inútil; 10–19 difícil captar la idea general; 20–29 la idea es clara, pero con errores gramaticales significativos; 30–40 traducciones comprensibles a buenas; 40–50 alta calidad; 50–60 muy alta calidad, adecuadas y fluidas; > 60 calidad frecuentemente superior a la humana. Consultado el 24 de agosto de 2026.
- Google Research (2026). Welcome to colab.research.google.com. Recuperado de <https://colab.research.google.com/>.
- Gow-Smith, E., Bérard, A., Boito, M. Z., and Calapodescu, I. (2023). Naver labs europe’s multilingual speech translation systems for the iwslt 2023 low-resource track.
- Groves, D. and Mundt, K. (2018). Friend or foe? google translate in language for academic purposes. *English for Specific Purposes*, 53:47–57.
- Hendy, A. et al. (2023). How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.
- Hernández, M. and Gómez, J. (2013). Aplicaciones de procesamiento de lenguaje natural. *Revista Politénica*, 32:87–96.
- Hernández-Sampieri, R. and Mendoza Torres, C. P. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill Education, México.
- Hornberger, N. H. and Coronel-Molina, S. M. (2004). Quechua language shift, maintenance, and revitalization in the andes: The case for language planning. pages 9–67. *International Journal of the Sociology of Language*.
- Jafari, Z. (2023). The role of ai in supporting indigenous languages. In *AI and Tech in Behavioral and Social Sciences*, pages 4–11.
- Joshi, P. et al. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of ACL*.
- Jurafsky, D. and Martin, J. H. (2019). Speech and language processing. pages 107–300. ICAC.

- Kano, T., Sakti, S., and Nakamura, S. (2021). Transformer-based direct speech-to-speech translation with transcoder. *IEEE Spoken Language Technology Workshop*, pages 958–965.
- Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.
- Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.
- Kłosowski, P. (2018). Deep learning for natural language processing and language modeling. *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 223–228.
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of Interspeech 2017*, pages 498–502.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. volume 26. NIPS.
- MinIO Inc. (2026). Minio object storage documentation. Consultado en abril de 2026.
- Ministerio de Educación (2013). Documento nacional de lenguas originarias (dnlo). Technical report, Ministerio de Educación del Perú, Lima, Perú.
- Ministerio de Educación del Perú (2018). Lenguas originarias del Perú. Technical report, Ministerio de Educación del Perú, Lima, Perú.
- Moreira, D., Cruz, I., Gonzalez, K., Quirumbay, A., Magallan, C., Guarda, T., Andrade, A., and Castillo, C. (2021). Análisis del estado actual de procesamiento de lenguaje natural. *Revista Ibérica de Sistemas e Tecnologías de Informação*, 42:126–136.
- Moreno Veliz, O. (2021). The repucs’ spanish–quechua submission to the americas nlp 2021 shared task on open machine translation. *Américas NLP 2021*.
- NLLB Team, Costa-jussà, M. R., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- OpenAI (2024). Gpt-4o system card. Disponible en <https://developers.openai.com/api/docs/models>.
- OpenAI (2026). Openai api platform documentation. Consultado el 17 de abril de 2026.
- Ortega, J. E., Mamani, R. C., and Cho, K. (2020). Traducción automática neuronal con un lenguaje polisintético. In *Machine Translation*.
- Ortega, J. E., Zevallos, R. J., Chen, W., and Abdulmumin, I. (2025). Quespa submission for the iwslt 2025 dialectal and low-resource speech translation task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 260–268.

- Paccotacya-Yanque, R. Y. G., Huanca-Anquise, C. A., Escalante-Calcina, J., Ramos-Lovón, W. R., and Cuno-Parari, Á. E. (2022). A speech corpus of quechua collao for automatic dimensional emotion recognition. *Scientific Data*, 9(1):778.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Popović, M. (2015). Chrf: character n-gram f-score for automatic mt evaluation. *In Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.
- PostgreSQL Global Development Group (2026). Postgresql documentation. Consultado en abril de 2026.
- Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Ramírez, S. (2024). Fastapi documentation. Consultado en abril de 2026.
- Raunak, V., Menezes, A., and Junczys-Dowmunt, M. (2021). The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). Comet: A neural framework for mt evaluation.
- Rousso, R., Cohen, E., Keshet, J., and Chodroff, E. (2024). Tradition or innovation: A comparison of modern asr methods for forced alignment. *arXiv preprint arXiv:2406.19363*.
- Sperber, M. and Paulik, M. (2020). Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421. Association for Computational Linguistics.
- Tanenbaum, A. S. and Van Steen, M. (2017). *Distributed Systems*. CreateSpace, 3rd edition.
- Tsiamas, I., Gállego, G., Fonollosa, J. A. R., and Costa-jussà, M. R. (2022). Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv preprint arXiv:2202.04774*.

- Vajjala, S., Majumder, B., and Gupta, A. (2020). Practical natural language processing: A comprehensive guide to building real-world nlp systems. In *Practical natural language processing: A comprehensive guide to building real-world NLP systems*, page 6.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, C. and Sennrich, R. (2020). On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552.
- Wu, H. et al. (2023). Using a forced aligner for prosody research. *Humanities and Social Sciences Communications*, 10:429.
- Xiangjie, K., Menglin, L., Kai, M., Kaiqi, T., Mengyuan, W., Zhaolong, N., and Feng, X. (2020). Revitalization of indigenous languages through preprocessing and neural machine translation: The case of inuktitut. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666.
- Zafar, M., Castaldo, A., Gajakos, N., Nayak, P., Haque, R., and Way, A. (2024). The setu-dcu submissions to iwslt 2024 low-resource speech-to-text translation tasks. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 80–85.