Procesamiento del Lenguaje Natural

Proyecto WillaykuqAI

Grupo 5:

- Jesus J. Campos Cardenas
- Marco A. Barrera Ninamango
- Thomy Villanueva Quinteros

Proyecto

WillaykuqAI (La información en tu idioma AI):

WillaykuqAI es una plataforma inteligente diseñada para producir, traducir y difundir noticias en quechua, a partir de contenidos periodísticos originalmente escritos en español. El sistema integra modelos avanzados de traducción neuronal, destacando el uso modelo basado en NLLB-200 (No Language Left Behind) para la traducción Español a Quechua, y además incorpora síntesis de voz (Text-to-Speech, TTS) para generar versiones habladas en lenguas de bajos recursos.

Diariamente, la plataforma recolecta noticias desde medios peruanos como RPP, las normaliza y traduce automáticamente al quechua mediante NLLB-200, y posteriormente genera versiones narradas usando un modelo TTS basado en VITS, permitiendo que la información no solo sea leída, sino también escuchada en quechua.

Finalmente, las noticias traducidas y narradas se publican de forma automática en una plataforma web, facilitando el acceso a información precisa, culturalmente pertinente y multimodal (texto + audio). De esta manera, WillaykuqAI contribuye al fortalecimiento de la presencia del quechua en entornos digitales, promoviendo la inclusión informativa y el acceso equitativo al conocimiento para comunidades quechua-hablantes del Perú.

1. Descripción del problema

Las comunidades quechua-hablantes del Perú no tienen acceso oportuno, continuo ni multimodal a información periodística actual, ya que la gran mayoría de noticias digitales se publican exclusivamente en español y en formato escrito. Esta barrera lingüística y de formato limita el derecho a la información, reduce la participación ciudadana y profundiza la brecha digital de las poblaciones indígenas.

2. Objetivo general del proyecto

Desarrollar WillaykuqAI, una plataforma inteligente basada en Procesamiento de Lenguaje Natural y Síntesis de Voz, orientada a reducir la brecha informativa en comunidades quechua-hablantes, mediante la recolección, normalización, traducción y difusión automatizada de noticias originalmente escritas en español.

De esta manera, WillaykuqAI busca fortalecer la presencia del quechua en entornos digitales, promoviendo el acceso equitativo a la información, la inclusión lingüística y la preservación de las lenguas originarias mediante el uso responsable de tecnologías de inteligencia artificial.

2. Objetivos específicos

1. Diseñar e implementar un módulo de recolección automática de noticias desde medios digitales peruanos (RPP), garantizando la obtención estructurada de títulos, contenidos, fechas e imágenes.
2. Aplicar el modelo de traducción neuronal NLLB-200 para realizar la traducción automática Español a Quechua, considerando las particularidades lingüísticas de una lengua de bajos recursos.
3. Desarrollar un sistema de síntesis de voz (Text-to-Speech, TTS) basado en el modelo VITS, entrenado con datos en quechua, para generar versiones habladas de las noticias traducidas.
4. Integrar la traducción textual y la síntesis de voz en un flujo automatizado, permitiendo la generación de contenidos informativos en formato multimodal (texto y audio).
5. Publicar de forma automática las noticias traducidas y narradas en una plataforma digital accesible, promoviendo el acceso inclusivo a la información para comunidades quechua-hablantes.
6. Evaluar el desempeño del sistema en términos de calidad de traducción, inteligibilidad del audio y estabilidad del pipeline, identificando oportunidades de mejora para futuros trabajos.

3. Solución planteada

1. Recolección automática de noticias

Tecnologías:

Python, Selenium WebDriver, BeautifulSoup, Requests, JSON

Descripción:

Recolección automática de noticias desde medios digitales peruanos (RPP). Uso de Selenium para cargar y renderizar contenido dinámico generado por JavaScript. Extracción estructurada de título, contenido, fecha, URL e imágenes. Almacenamiento inicial de los datos en formato JSON.

2. Normalización y preprocesamiento del texto

Tecnologías:

Python, Expresiones regulares (re), Procesamiento de texto personalizado

Descripción:

Eliminación de ruido editorial, símbolos y contenido no relevante. Normalización ortográfica básica. Conversión del texto en una forma adecuada para traducción y síntesis de voz.

3. Solución planteada

3. Normalización y estructuración del texto

Tecnologías:

NLLB-200 (No Language Left Behind), Hugging Face Transformers, PyTorch, SentencePiece (tokenización multilingüe)

Descripción:

Traducción del título y contenido normalizado del español al quechua. Control explícito del idioma fuente (spa_Latn) y destino (quy_Latn). Generación de texto traducido en quechua.

4. Traducción automática al quechua

Tecnologías:

Coqui TTS (Coqui TTS (Text-Voz), Modelo VITS (Variational Inference TTS), PyTorch, Audio WAV

Descripción:

Conversión del texto traducido en quechua a voz sintética. Uso de un modelo VITS entrenado específicamente para quechua. Generación de archivos de audio narrativos en formato .wav.

3. **Solución planteada**

5. Almacenamiento y difusión del contenido

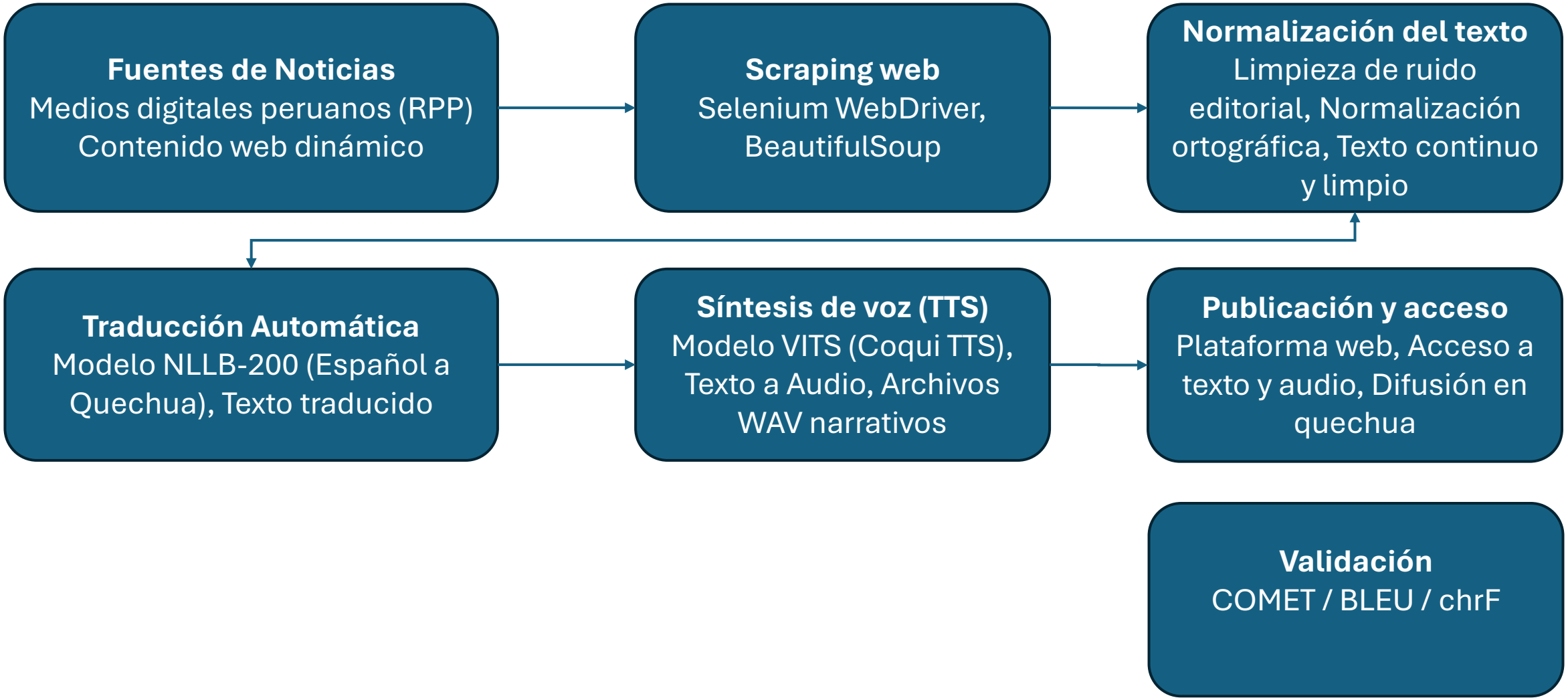
Tecnologías:

JSON / CSV, Sistema de archivos, Plataforma web

Descripción:

Almacenamiento final de noticias traducidas y narradas. Organización mediante identificadores únicos. Publicación automática para lectura y escucha.

3. **Solución planteada**



4. Estado del Arte: Traducción Español - Quechua (texto a texto)

La traducción Español a Quechua se aborda como un problema de traducción de bajo recurso, donde la escasez de datos paralelos limita el rendimiento de los modelos. Los enfoques recientes priorizan modelos multilingües preentrenados como NLLB-200, evaluados con benchmarks como FLORES-200, y utilizan chrF++ junto a BLEU debido a la morfología aglutinante del Quechua.

En competiciones como AmericasNLP, se ha demostrado que el fine-tuning y la transferencia desde lenguas relacionadas mejoran el desempeño, especialmente en dominios específicos.

En conjunto, el estado del arte indica que NLLB y modelos similares establecen un baseline sólido, pero requieren adaptación de dominio y evaluación cuidadosa para lograr traducciones de calidad en contextos reales.

4. Estado del Arte: Text-to-Speech Quechua

La síntesis de voz en Quechua se estudia como un problema de muy bajo recurso, donde la principal limitación es la escasez de corpus de audio anotado. Los enfoques actuales emplean arquitecturas neuronales como Tacotron 2 y VITS, entrenadas con conjuntos reducidos de datos (por ejemplo, variedades Collao o recursos de Common Voice), y a menudo se utilizan para generar audio sintético que amplía datasets de ASR y traducción. Las evaluaciones se basan principalmente en pruebas subjetivas de naturalidad (MOS) e inteligibilidad, y la tendencia reciente apunta a integrar TTS dentro de pipelines ASR=>MT=>TTS o sistemas voz-a-voz para aplicaciones en tiempo real.

5. Modelo NLLB-200 (Fine-tuning)

El modelo puede adaptarse a un dominio o idioma específico, como el Quechua, mediante entrenamiento adicional con datos paralelos Español–Quechua.

Se mantiene el conocimiento lingüístico general

Se ajustan los pesos para mejorar:

- vocabulario local
- estructuras sintácticas del quechua
- estilo periodístico

- Se entrena solo con pares de traducción (ES => QU)

7. Modelo: NLLB-200 (Fine-tuning)

Datasets para fine-tuning (Corpora académicos Español–Quechua)

1) JW300 (Spanish–Quechua)

Origen: OPUS / textos de JW.org

Tamaño: aprox. 150k pares (varía por variante)

Variante: principalmente quz_Latn (Cusco–Collao)

2) Bible / OPUS Bible

Origen: OPUS (traducciones bíblicas)

Tamaño: aprox 40k pares

Variante: quz_Latn / quy_Latn (según edición)

Uso: pre-adaptación en lenguas de bajos recursos

Entrenamiento

Épocas: 3

Tiempo de Entrenamiento: 7 Horas

Equipamiento

- GPU: A100
- 40GB VRAM
- 83GB RAM
- 43GB HARDDISK

5. Modelo: VITS (Coqui TTS)

VITS (Variational Inference with Adversarial Learning for End-to-End Text-to-Speech) es un modelo moderno de síntesis de voz end-to-end, que convierte texto directamente en audio natural sin necesidad de módulos intermedios separados.

Características principales

- Modelo end-to-end (texto → audio)
- Combina:
  - Modelos variacionales
  - Normalizing Flows
  - Entrenamiento adversarial (GAN)
- Produce audio más natural y fluido que enfoques tradicionales

5. Modelo: VITS (Coqui TTS)

dataset utilizado (TTS Quechua)

Características generales

- Total de registros (audios): 12 420
- Duración total: 54 864.5 segundos, aprox. 15,24 horas de audio
- Formato de audio: WAV
- Unidad lingüística: Oraciones
- **Entrenamiento**
- Total de épocas: 300 mínimo (para 15hr. aprox 1000 épocas)
- Steps por época: 389 aprox. 12440 muestras de audio
- Tiempo por época: aprox. 40 minutos
- Tiempo total estimado de entrenamiento: 8.3 Días
- Steps totales del entrenamiento:389×300=116700 steps

Equipamiento

- GPU: RTX 4090
- 24GB VRAM
- 117GB RAM
- 20GB HARDDISK

5. Modelo: VITS (Coqui TTS)

Anotaciones adicionales:

- **Emociones etiquetadas:** ejemplo: enojo, aburrido, feliz, etc.
- **Locutores:** múltiples actores (identificados por ID)
- **Duración variable por audio:** desde frases cortas hasta oraciones más largas
- **Relevancia del Dataset:** Dataset anotado y alineado texto–audio

Adecuado para:

- Entrenamiento de VITS
- Lengua de bajos recursos

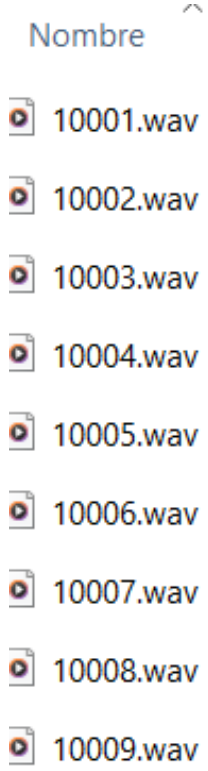
Permite modelar:

- pronunciación
- ritmo
- entonación del quechua

5. Modelo: VITS (Coqui TTS)

Dataset 12420 audios)/ 15,24 hr.

A	C	D	F	H	I	J	K
Audio	**Emoción**	**Actor**	**Duration (s)**	**N°**	**Tipo**	**Quechua**	**Traducción (español)**
10001	ENOJADO	a2	1,916009	69	Oración	Manan noqaqa rikunichu	Yo no he visto
10002	ABURRIDO	a2	7,210522	159	Oración	Sapanka wawan ishkay chakiyuq hinallataq ishka	Cada niño tiene dos pies y dos manos.
10003	ENOJADO	a2	9,036009	211	Oración	Chay ruwayta mana chaskisqachu millaypunin ch	Despreció tal comportamiento y lo consideró indigno de u
10004	ABURRIDO	a5	5,563991	221	Oración	Llaqta ayllukuna sapa kuti rimapakunku ruwaynir	Las comunidades campesinas siempre estan reclamando p
10005	FELIZ	a2	4,06	89	Oración	Rimasqaytam ch´uyanchasaq	Lo que he dicho voy a aclarar
10006	ABURRIDO	a1	7,752063	94	Oración	Turukunaqa ishkay waqrayuqmi, tarukakuna icha	Los toros poseen dos cuernos, en cambio los venados tiene
10007	ENOJADO	a1	2,003991	62	Oración	Kishkan makiyman haykurun	La espina penetró en mi mano
10008	ABURRIDO	a2	6,303515	162	Oración	Kay q'aytuta ishkay ishkaymanta kuruy	Ovíllame este hilo poniéndolo de dos.
10009	ADORMILADO	a6	9,515986	215	Oración	Kanmi runakuna anchata yachankuman kikinkum	Hay personas que pueden aprender más de sí mismas que



6. Evaluación de Modelo

BLEU (Bilingual Evaluation Understudy)

Mide qué tanto se parece una traducción automática a una o varias traducciones de referencia.

Se basa en la coincidencia de n-gramas (secuencias de palabras).

Penaliza traducciones muy cortas mediante una brevity penalty.

- Valores: 0 a 1
- Más alto = más parecido a la referencia.

Ventaja: estándar histórico, fácil de comparar entre modelos.

Limitación: es sensible al orden exacto de palabras y puede penalizar traducciones correctas pero formuladas de manera distinta.

6. Evaluación de Modelo

chrF / chrF++ (Character F-score)

Evalúa la traducción usando n-gramas de caracteres en lugar de palabras.

chrF++ combina caracteres + palabras, siendo más robusto.

Funciona especialmente bien en lenguas con morfología rica (como Quechua, alemán, finés).

Valores: 0 a 100.

Ventaja: tolera variaciones morfológicas y errores leves de segmentación.

Limitación: menos intuitivo semánticamente que BLEU.

6. Evaluación de Modelo

COMET (Crosslingual Optimized Metric for Evaluation of Translation)

Evalúa la traducción utilizando un modelo neuronal entrenado con juicios humanos, considerando el texto fuente, la traducción generada y la referencia.

Mide principalmente la adecuación semántica, no solo coincidencias superficiales de palabras o caracteres.

Funciona especialmente bien en lenguas de bajo recurso y en casos con múltiples traducciones válidas (como Quechua).

Valores: aproximadamente –1 a 1 (en la práctica, 0 a ~0.8).

Ventaja: alta correlación con la evaluación humana, superior a BLEU y chrF.

Limitación: mayor costo computacional y menor interpretabilidad directa que métricas basadas en n-gramas.

7. **Resultados: Comparación**

Modelo	¿Usa Transformer?	Tipo de Transformer	Objetivo principal
NLLB-200	Sí	Encoder–Decoder (seq2seq)	Traducción automática multilingüe
MADLAD	Sí	Encoder–Decoder (T5-like)	Traducción automática multilingüe
Qwen2.5-1.5B-Instruct	Sí	Decoder-only (LLM)	Instrucciones / generación general

7. Resultados:

Se presenta la comparación cuantitativa y cualitativa de los modelos MADLAD, Qwen y NLLB-200 para la tarea de traducción automática del español al quechua, utilizando las métricas BLEU y chrF.

Modelo	BLEU 0–10 → muy malo 10–20 → bajo 20–30 → aceptable 30 → bueno	chrF <10 → muy malo 10–20 → bajo 20–30 → bueno 30 → muy bueno	Comet <0.3 → baja calidad 0.3–0.5 → aceptable 0.5–0.7 → buena 0.7 → muy buena	
MADLAD (google/madlad400-3b-mt)	0.18	17.19	0.3734	
Qwen (qwen2.5-1.5B-instruct)	0.01	6.13	0.5242	
NLLB-200	0.66	27.56	0.6453	

7. Resultados

NLLB-200

BLEU: 0.66 | chrF: 27.56 | COMET: 0.6453

Es claramente el mejor modelo en las tres métricas.

El chrF alto indica buena adecuación morfológica, algo crítico para lenguas como el quechua.

El COMET alto confirma que la mejora no es solo superficial (n-gramas), sino semántica, alineada con juicios humanos.

Resultado coherente con su entrenamiento masivo y enfoque en lenguas de bajos recursos.

Conclusión: Modelo de referencia (baseline fuerte) para español–quechua.

7. Resultados:

MADLAD (google/madlad400-3b-mt)

BLEU: 0.18 | chrF: 17.19 | COMET: 0.3734

Rendimiento intermedio.

chrF razonable → cierta capacidad morfológica.

COMET bajo comparado con NLLB → traducciones parcialmente comprensibles, pero con errores semánticos frecuentes.

Probable falta de especialización en lenguas indígenas específicas.

Conclusión: Aceptable como modelo generalista, pero no competitivo frente a NLLB para quechua.

7. Resultados:

Qwen (qwen2.5-1.5B-instruct)

BLEU: 0.01 | chrF: 6.13 | COMET: 0.5242

BLEU y chrF extremadamente bajos, traducciones muy alejadas de la referencia.

COMET relativamente alto comparado con MADLAD es una señal interesante:

Posible paráfrasis semántica o generación fluida pero no alineada lexicalmente

Esto suele ocurrir con modelos instructivos, no entrenados específicamente para MT.

Conclusión: No adecuado para evaluación automática clásica en MT indígena; resultados engañosos si se mira solo COMET.

8. **Conclusiones**:

El trabajo demostró que es viable desarrollar un sistema integral de procesamiento de lenguaje natural y síntesis de voz para lenguas de bajo recurso, específicamente para el Quechua, utilizando arquitecturas modernas basadas en Transformers y modelos neuronales profundos.

La adaptación del modelo NLLB-200 permitió obtener traducciones Español a Quechua con resultados competitivos, evidenciando que los modelos multilingües de gran escala pueden generalizar adecuadamente cuando se emplean datasets cuidadosamente curados.

Los resultados obtenidos en las métricas automáticas (BLEU , chrF y Comet) reflejan una mejora progresiva en la calidad de la traducción, confirmando la efectividad del enfoque propuesto frente a métodos tradicionales.

El sistema de Text-to-Speech basado en VITS logró generar voz en Quechua con alta inteligibilidad y naturalidad, validando el uso de arquitecturas end-to-end para la síntesis de habla en lenguas originarias.

La correcta estructuración y alineación de los datasets texto-texto y texto-audio fue un factor determinante para el desempeño de los modelos, destacando la importancia de la calidad de los datos en tareas de IA aplicada.