PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Síntesis de voz para lenguas de pocos recursos: El caso Shipibo-Konibo

Trabajo de investigación para obtener el grado académico de Maestro en Informática con mención en Ciencias de la Computación que presenta:

Daniel Arturo Menéndez Quinto

Asesor:

Héctor Erasmo Gómez Montoya

Lima, 2024

Informe de Similitud

Yo, Héctor Erasmo Gómez Montoya, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor(a) de el trabajo de investigación titulada(o) Síntesis de voz para lenguas de pocos recursos: El caso Shipibo-Konibo, de el autor Daniel Arturo Menéndez Quinto, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 4%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 10 de Diciembre del 2024.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de investigación, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

Lima, 12 de Diciembre del 2024.

Apellidos y nombres del asesor / de la asesora: Gómez Montoya, Héctor Erasmo	
DNI: 70599170	Firma
ORCID: 0000-0002-1338-3392	

Agradecimientos

En primer lugar, quiero expresar mi más sincero agradecimiento a la comunidad Shipibo-Konibo por medio de Seleni Rojas (Inkan Jabe) una activista por la conservación de su fascinante cultura y sin quien no hubiera sido posible recabar los datos necesarios para el desarrollo de este trabajo.

También quiero agradecer a la Universidad Nacional Intercultural de la Amazonía y al grupo de investigación Chana para las ciencias del lenguaje y la interculturalidad de la Pontificia Universidad Católica del Perú por el invaluable apoyo durante las pruebas del modelo mostrado en este artículo. Gracias también a personas como los profesores Roberto Zariquiey y Jorge Sato que dieron su apoyo y consejo durante el desarrollo del modelo de síntesis de voz.

Resumen

Actualmente, existe consenso entre numerosos lingüistas en que, de las más de 7000 lenguas conocidas en el mundo, muchas están en peligro de extinción en distintos grados. Por ello, su documentación y revitalización son tareas esenciales, no solo para conservarlas, sino también para preservar formas únicas de comunicación y valiosas maneras de comprender el mundo.

Esta investigación busca ser un paso inicial en la revitalización de lenguas amenazadas, enfocándose en el Shipibo-Konibo. Esta lengua, hablada principalmente en la Amazonía peruana, enfrenta desafíos como la escasez de datos, la coexistencia de diferentes tradiciones ortográficas y una documentación limitada, lo que lo clasifica como una lengua de pocos recursos.

En respuesta a estos retos, este artículo presenta el desarrollo de un modelo de síntesis de texto a voz (TTS) para el Shipibo-Konibo basado en la arquitectura Tacotron 2 y HiFi-GAN como vocoder, superando diversas dificultades técnicas para lograr una solución capaz de generar audio de alta calidad.

Se requirió la recopilación de un corpus que incluye más de 4 horas de grabaciones y 3,025 frases escritas, obtenidas de textos educativos y traducciones literarias. Las grabaciones fueron realizadas con la ayuda de un hablante nativo, asegurando un alto estándar de calidad para el entrenamiento del modelo.

Los resultados fueron prometedores, alcanzando una tasa de inteligibilidad del 88.56% y una puntuación media de opinión (MOS) de 4.01. Estas métricas llegaron incluso a superar la calidad percibida de la voz natural en las pruebas realizadas, lo que demuestra el potencial del modelo para adaptarse a otros idiomas de la familia pano u otras lenguas amazónicas.

ÍNDICE

LISTA DE TABLAS

LISTA DE FIGURAS

Síntesis de voz para lenguas de pocos recursos: El caso del Shipibo-Konibo

Daniel Menéndez
Escuela de Posgrado
Pontificia Universidad Católica del Perú
dmenendez@pucp.edu.pe

Héctor Erasmo Gómez
Grupo de Investigación Chana
Escuela de Posgrado
Pontificia Universidad Católica del Perú
hector.gomez@pucp.edu.pe

Abstract

Este artículo presenta el diseño y desarrollo de un modelo de Texto a Voz (TTS) para el Shipibo-Konibo, una lengua indígena de pocos recursos hablada principalmente en la Amazonía del Perú. A pesar del desafío que presentó la escasez de datos, el modelo fue entrenado con más de 4 horas de grabaciones y 3025 frases escritas recopiladas meticulosamente. Los resultados mostraron una tasa de inteligibilidad (IR) superior al 88 % y una calificación media de opinión (MOS) de 4.01, lo que confirma la calidad del audio generado por el modelo, compuesto por el predictor de espectrogramas Tacotron 2 y el vocoder HiFi-GAN. Además, se destaca el potencial de este modelo para entrenarse en otros idiomas indígenas hablados en Perú, abriendo un camino prometedor para la documentación y revitalización de estas lenguas.

1 Introducción

Alrededor del mundo existen aproximadamente 7169 idiomas (SIL, Accessed March 14 2024) y numerosos lingüistas han reconocido cada vez más la necesidad imperativa de documentar y revitalizar muchos de estos idiomas, ya que corren peligro de desaparecer(Krauss, 1992; Wurm, 1956; Zaborski, 1970; Capell, 1962; Becker-Donner, 1962; Stone, 1962; Kibrik, 1991; Wurm, 1991; Adelaar, 1991; Swadesh, 1960; Krauss, 2007; Sands, 2017; Campbell and Rehg, 2018). La pérdida de diversidad lingüística representa una amenaza significativa para nuestra comprensión de los distintos sistemas de comunicación de nuestra especie, así como su evolución y diseminación entre varios grupos étnicos en diferentes regiones del mundo (Evans and Levinson, 2009). Además, conlleva la erosión global de un conocimiento tradicional invaluable, esencial para el sustento de los pueblos indígenas en todo el mundo (Cámara-Leret and Bascompte, 2021).

En el contexto particular del Perú, el conteo preciso de los idiomas originarios hablados en este país sigue siendo objeto de debate. Mientras que las estadísticas oficiales reconocen 48 idiomas (4 en los Andes y 44 en la Amazonía), junto con una lengua de señas vernácular (SLP), Glottolog enumera 90 idiomas existentes en esta nación(Hammarström et al., 2021). Es sabido que una parte considerable de estos idiomas enfrenta diversos grados de peligro de desaparición y solo la mitad de ellos posee algún nivel de documentación contemporánea confiable. Entre estos idiomas se encuentra el Shipibo-Konibo, perteneciente a la familia lingüística Pano. Con aproximadamente 40,000 hablantes[1], es el idioma más vibrante dentro de esta familia. Es mayormente hablado en las regiones peruanas de Loreto, a lo largo del río Ucayali y sus afluentes (Valenzuela, 2003). Otras comunidades Shipibo-Konibo también se localizan en Lima, Huánuco y Madre de Dios.

La formalización del alfabeto Shipibo-Konibo por parte del Ministerio de Educación del Perú en el año 2015, que consta de 19 caracteres (BDPI-MINCUL, 2018), marcó un hito significativo en el proceso de normalización del idioma. Antes de este evento los materiales educativos, libros y documentación en Shipibo-Konibo utilizaban un sistema ortográfico distinto, desarrollado en la década de 1960 por misioneros de SIL. En consecuencia, la coexistencia actual de al menos dos tradiciones ortográficas para el idioma representa un gran desafío en lo que respecta al uso de materiales escritos en esta lengua para el desarrollo de algoritmos de aprendizaje profundo y proyectos de aprendizaje automático.

En este contexto, el presente artículo presenta los esfuerzos realizados para desarrollar un modelo inicial de síntesis de texto a voz para el idioma Shipibo-Konibo, una tarea crucial en el ámbito del procesamiento del lenguaje natural (NLP). La fase

[1]El censo peruano de 2017 estima la población total de Shipibo-Konibo en 34,000, pero se espera que la cifra real sea más alta (INEI, 2018)

Figura 1: Mapa del Perú que muestra las zonas donde se ubican las diversas comunidades Shipibo-Konibo.

inicial consistió en identificar los desafíos inherentes al desarrollo de modelos TTS para idiomas con recursos limitados, le siguió la selección de un modelo neuronal óptimo y la proposición de estrategias para superar los desafíos mencionados. Posteriormente, se recolectaron y procesaron minuciosamente datos de texto y audio para crear un conjunto de datos apropiado para el aprendizaje profundo. Luego se determinaron los recursos computacionales necesarios para entrenar y evaluar el modelo TTS propuesto. Finalmente, se realizaron pruebas y se validaron los resultados. A lo largo de este documento, se detalla este proceso integral, con el deseo de que sirva de inspiración para futuros proyectos de modelos TTS dirigidos a otras lenguas indígenas del Perú y América Latina. Asimismo, se espera que el modelo TTS desarrollado tenga el potencial de adaptarse con éxito a otros idiomas de la familia Pano.

2 Síntesis de voz para lenguas de pocos recursos

El proceso de desarrollar un modelo de síntesis de texto a voz para idiomas con pocos recursos como el Shipibo-Konibo presentó una serie de desafíos. Principalmente, en el contexto del aprendizaje profundo, se tuvo que afrontar el reto de la adquisición de una cantidad suficiente de datos estructurados para entrenar efectivamente un modelo. Este obstáculo inicial condujo a un desafío posterior por el cual se condicionó el establecimiento de estrategias de entrenamiento como la transferencia de aprendizaje y la utilización de arquitecturas neuronales especializadas para lograr un rendimiento óptimo del modelo con el fin de producir resultados de voz claras y naturales.

Desafíos adicionales surgieron al estimar los recursos computacionales necesarios y seleccionar la plataforma más adecuada para entrenar el modelo TTS. Por otro lado, la evaluación subjetiva de la voz sintética implico la aplicación de métricas apropiadas y adaptadas a los matices de los idiomas con recursos limitados. Finalmente, tras analizar los datos, se obtuvieron resultados prometedores.

3 El modelo TTS propuesto para el Shipibo-Konibo

El reciente avance en redes neuronales profundas (DNNs) ha facilitado el desarrollo de varios modelos de texto a voz en los últimos años, incluyendo ejemplos como: Transformer-TTS (Li et al., 2019), Glow-TTS (Kim et al., 2020), Deep Voice (Arık et al., 2017), y Tacotron (Arık et al., 2017). Varios de estos han demostrado un rendimiento impresionante por lo que se establecieron criterios para la selección del modelo con el fin de identificar la opción más adecuada para el idioma Shipibo-Konibo.

Inicialmente, se descartaron modelos que han caido en desuso como Voice Loop 2 (Taigman et al., 2017). Posteriormente, se considero acudir a modelos previamente utilizados en experimentos similares (Gopalakrishnan et al., 2022), lo que acortó las opciones viables. Luego se indagó por investigaciones anteriores realizadas en entornos de pocos recursos y se obtuvieron alternativas como Deep Voice (Ping et al., 2017), Tacotron 2 (Shen et al., 2018), y FastSpeech2 (Ren et al., 2020).

Finalmente, Tacotron 2 fue seleccionado como el modelo más apropiado. Esta decisión se basó en resultados obtenidos en estudios similares realizados en la India, donde Tacotron 2 se utilizó con éxito en un idioma de recursos limitados como el sánscrito (Debnath et al., 2020), así como en una experiencia que involucro un idioma tribal con recursos mínimos, como el Lambani (Dasare et al., 2022), donde Tacotron 2 produjo resultados prome-

tedores con un conjunto de datos de datos de solo 1000 frases.

3.1 Tacotron 2

Tacotron 2 (Shen et al., 2018) es un modelo de síntesis de voz desarrollado por Google e implementado por Nvidia, que presenta una arquitectura de codificador-atención-decodificador. Su característica más resaltante es que incorpora mecanismos de «atención sensible a la ubicación» para mejorar la calidad final del habla sintetizada. El codificador, que constituye el componente inicial, traduce la secuencia de caracteres en un embedding, que posteriormente sirve como entrada al decodificador para predecir los espectrogramas Mel, que son mejores para la representación de voz humana.

Para este proyecto, se empleó la implementación de Tacotron 2 en PyTorch, y el entrenamiento se realizó utilizando técnicas de transferencia de aprendizaje de acuerdo con los parámetros del conjunto de datos descritos en la Sección 4.

3.2 El Vocoder Hi-Fi GAN

A diferencia del enfoque descrito en el artículo original de Tacotron 2 donde se proponía usar el vocoder WaveGlow (Prenger et al., 2019) , este estudio ha optado por utilizar el modelo de síntesis de voz HiFi-GAN (Kong et al., 2020). Esta decisión surge de experiencias recientes de desarrolladores con tareas similares, donde este vocoder ha demostrado un rendimiento superior en comparación con WaveGlow.

HiFi-GAN opera como una red generativa adversarial (GAN), que consta de un generador y dos discriminadores (multi-escala y multi-periodo), junto con dos funciones de pérdida asociadas. Este vocoder es capaz de generar formas de onda en el dominio del tiempo (audio) a partir de los espectrogramas predichos por el modelo Tacotron 2 como se puede ver en la Figura 2.

4 Recopilación de Datos de Voz y Texto para el Entrenamiento

El proceso de recopilación de datos para el entrenamiento del modelo planteó varios desafíos. La introducción de un nuevo alfabeto en 2015 hizo que los recursos escritos previamente disponibles, como las traducciones de la Biblia y otras obras literarias, quedaran parcialmente obsoletos. Por lo que se volvió imperativo buscar nuevos recursos.

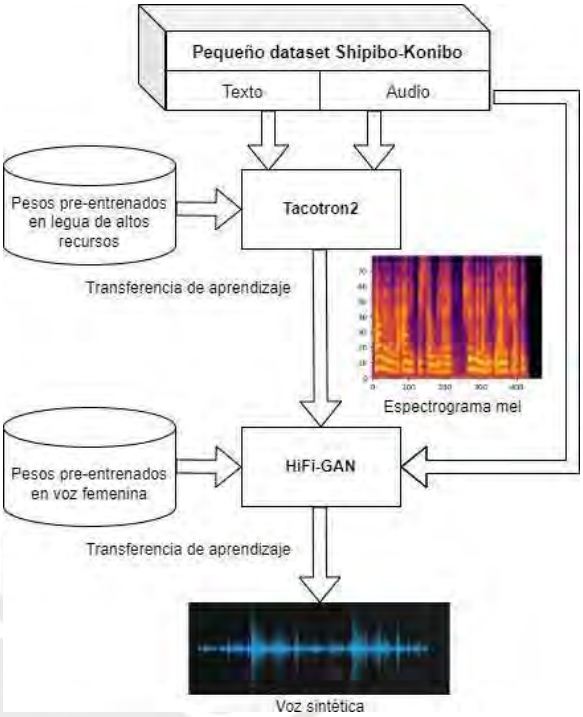

Figura 2: Diagrama de bloques del modelo TTS propuesto.

4.1 Recopilación de Texto

Encontrar textos adecuados resultó ser un gran desafío. En primer lugar, gracias a una publicación de la asociación cultural Alianza Francesa, se consiguió la traducción al Shipibo-Konibo del cuento «El Principito» («Jatibi Ibo Bake») del 2018. De esta fuente se extrajeron aproximadamente 1200 frases para incluirlas en el conjunto de datos.

Posteriormente, fue posible descargar desde el Ministerio de Educación del Perú textos utilizados en la educación bilingüe desde el año 2017. Esta colección abarcaba guías de enseñanza, manuales instructivos y cuentos infantiles, lo que culminó en la recolección de siete publicaciones. A partir de estos materiales, se extrajeron más de 2000 frases. En última instancia, mediante la fusión de ambas fuentes, se compiló un corpus que comprendía un total de 3025 frases.

4.2 Recopilación de Audio

Los requisitos para el conjunto de datos de audio de entrenamiento que el modelo Tacotron 2 requiere son bastante específicos:

- **F**recuencia de muestreo de 22050Hz o superior.
- **U**n solo orador.

- **L**a frases deben tener fonemas diversos.
- **L**a duración de los audios debe estar entre 1 a 10 segundos.
- **L**os segmentos de audio no deben tener silencios al principio ni al final.
- **L**os segmentos de audio no deben contener pausas largas.

La comunidad de la Pontificia Universidad Católica del Perú (PUCP) cuenta con una notable diversidad cultural, y entre sus miembros, una joven contribuyó con entusiasmo al desarrollo de este proyecto. Como hablante nativa y activista de la conservación del Shipibo-Konibo, aportó ideas invaluables y experiencia al esfuerzo. Afortunadamente, tenía experiencia previa en la realización de grabaciones para la documentación de su lengua materna.

Las sesiones de grabación se llevaron a cabo durante cuatro tres meses, sesiones organizadas estratégicamente en intervalos de no más de 2 horas para mitigar la fatiga vocal y mantener una calidad consistente a lo largo del proceso.

Al terminar todas las sesiones, el conjunto de datos final presentaba las características mostradas en la Tabla 1.

Caracteristicas	Valor
Número de frases	3025
Duración total	$4h37m14s$
Mínima duración de frase	$1{,}08s$
Máxima duración de frase	$12{,}1s$
Duración promedio de frase	$7{,}4s$

Tabla 1: Características del conjunto de datos de audio recolectado de la lengua Shipibo-Konibo.

Finalmente, las frases fueron recortadas, remuestreadas de 44.1KHz a 22.05KHz, se eliminaron los silencios que superaran los 0.5 segundos que no correspondíeran a signos de puntuación y se normalizaron el volumen, la velocidad y el texto.

5 Experimentación y Resultados

Es crucial enfatizar que los modelos Tacotron 2 y HiFi-GAN requirieron entrenamientos separados y secuenciales debido a la dependencia del proceso de entrenamiento del vocoder en los resultados del modelo de predicción de espectrogramas, como se muestra en la Figura 2. Además, aprovechando las experiencias previas con lenguas y dialectos indios (Debnath et al., 2020; Dasare et al., 2022), el ajuste fino de los hiperparámetros para este proyecto resultó ser más rápido y efectivo.

5.1 Entrenamiento del Modelo Tacotron 2

Gracias a la investigación relizada sobre experiencias similares (Gopalakrishnan et al., 2022), la plataforma Google Colab surgió como la opción más viable, donde se dispuso de una GPU Nvidia A100 recomendada para el mejor entrenamiento de modelos de lenguaje.

Al adoptar la estrategia de transferencia de aprendizaje, es importante mencionar a la comunidad de Hugging Face (Pantoja, 2023) por facilitarar un modelo preentrenado en español latinoamericano con voz femenina. Esta selección se basó en la proximidad geográfica entre el español y el Shipibo-Konibo, lo que resultó ser una decisión acertada, como lo demuestra el proceso de entrenamiento, que mostró un alineamiento bastante rápido.

Las Tablas 2 y 3 presentan los hiperparámetros y los parámetros del optimizador Adam, respectivamente, mostrando las configuraciones que produjeron un rendimiento óptimo.

Hiperparámetros	Valor
Épocas	225
Tamaño de lote	16
Umbral del gate	0,5
Dropout de decodificador	0,1
Atención	0,1

Tabla 2: Hiperparámetros del modelo Tacotron2.

Parámetros del optimizador	Valor
Tasa de aprendizaje	$3{,}10^{-4}$
β_1	0,9
β_2	0,999
Weight decay	$1{,}10^{-5}$

Tabla 3: Parámetros del optimizador Adam.

Como se mencionó, el alineamiento entre el codificador y el decodificador mostró resultados prometedores desde la primera época, como se ilustra en la Figura 3.

El progreso del entrenamiento fue monitoreado durante todo el proceso, revelando mejoras consistentes tanto en el gráfico de alineamiento como en la pérdida de validación, como se muestra en la Figura 4. Sin embargo, alrededor de la 225ª época, su evolución se estancó, con una pérdida de validación

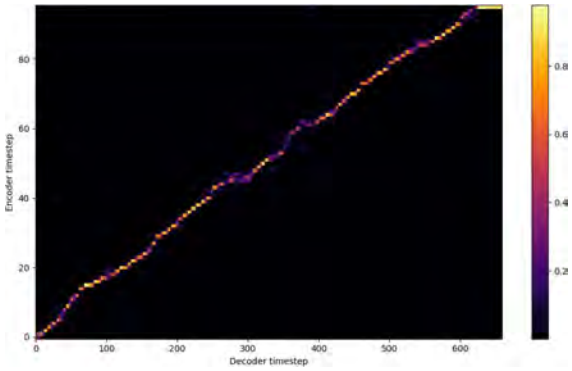

Figura 3: Alineamiento en la primera época.

estabilizándose en aproximadamente 0.143432. En consecuencia, se tomó la decisión de terminar el entrenamiento, habiendo logrado la alineamiento mostrado en la figura 5.

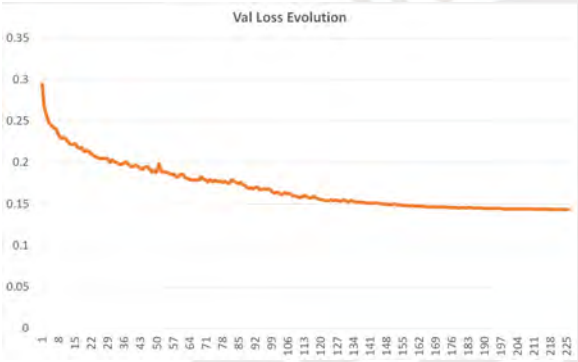

Figura 4: Evolución de la función de pérdida en validación.

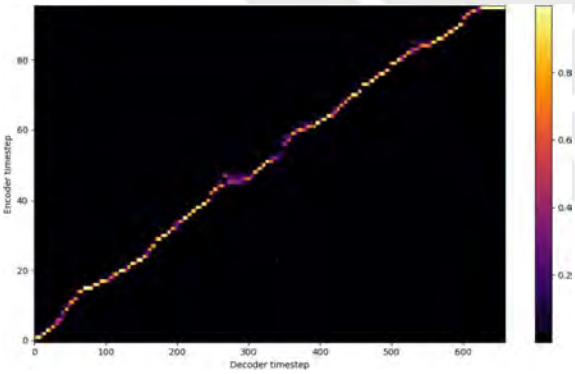

Figura 5: Alineamiento final en la época 225.

5.2 Entrenamiento del vocoder HiFi-GAN

Para entrenar el vocoder, fue esencial utilizar el modelo Tacotron 2 entrenado en la primera fase. Este modelo es responsable de generar espectrogramas, que posteriormente son convertidos en audio por el vocoder, como se ilustra en la Figura 2. Las muestras de audio originales del conjunto de datos recopilado se utilizaron con fines de comparación junto con el audio sintético.

Como estrategia de transferencia de aprendizaje, se recurrió a un modelo universal preentrenado con voz femenina. Después de 34 épocas (aproximadamente 5000 pasos), se hizo evidente que la función de pérdida había alcanzado una meseta y dejó de evolucionar.

5.3 Evaluación y resultados

Con el fin de recopilar las frases requeridas para la evaluación del modelo, el grupo Chana para las Ciencias del Lenguaje y la Interculturalidad de la PUCP facilitó un libro publicado antes del 2015 pero con un alfabeto facil de adaptar al vigente. El libro se titula *«Koshi Shinanya Ainbo»* («El Testimonio de una Mujer Shipiba») (Ainbo, 2005), que fue originalmente escrito en el idioma Shipibo-Konibo y narra historias sobre la vida y tradiciones de la Sra. Ranin Ama y su comunidad.

Como prueba inicial de inferencia, la figura 6 muestra el resultado para una frase extraida del libro antes mencionado, consta de cinco palabras y dura tres segundos: *«Nokon titan ea axeani jawékibo»*, que se traduce como «Las cosas que mi madre me enseño».

Establecer un protocolo de pruebas para modelos de lenguaje de bajos recursos fue un desafío significativo. Las pruebas objetivas, como el PESQ (Rix, 2003) o POLQA (Beerends et al., 2013), requieren acceso a abundantes datos preexistentes de audio y texto de gran calidad que no están disponibles en el caso del Shipibo-Konibo.

Para superar estas limitaciones, se recurrió a pruebas subjetivas, que aunque demandan más tiempo y esfuerzo, ofrecen una evaluación inicial confiable de la calidad del modelo. Entre las pruebas subjetivas comúnmente utilizadas para evaluar modelos de síntesis de texto a voz se encuentran la tasa de inteligibilidad (IR) y la calificación media de opinión (MOS).

La tasa de inteligibilidad (IR) evalúa la claridad y comprensibilidad del audio sintetizado para los oyentes. Se presentan muestras de audio sintetizadas a un grupo de evaluadores humanos, quienes intentan transcribirlas palabra por palabra. La tasa de inteligibilidad se calcula como el porcentaje de palabras correctamente identificadas o transcritas en comparación con el texto original.

Por otro lado, la calificación media de opinión

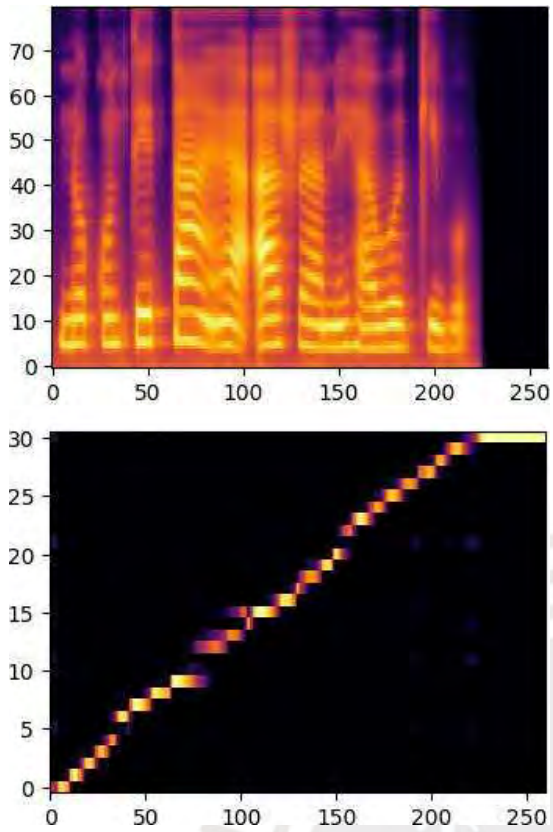

Figura 6: Espectrograma y gráfico de alineamiento del archivo WAV de la frase de 3 segundos.

(MOS) se basa en las puntuaciones de un grupo de evaluadores que escuchan muestras de audio generadas por el modelo y evalúan su calidad puntuando la claridad, naturalidad y fluidez percibida en una escala numérica de 1 a 5, donde:

Valor	Descripción
1	Calidad inaceptable o muy pobre
2	Calidad deficiente
3	Calidad aceptable o adecuada
4	Buena calidad
5	Excelente calidad

Tabla 4: Escala de la métrica MOS.

El valor promedio de las calificaciones otorgadas por los evaluadores se utiliza para determinar la calificación final de media de opinión.

Con las métricas definidas, se acudió a 26 evaluadores nativos de la lengua Shipibo-Konibo. 11 hombres y 15 mujeres menores de 30 años, la mayoría de ellos provenientes de comunidades Shipibo-Konibo de la zona del alto Ucayali y familiarizados con el uso de herramientas tecnológicas (WhatsApp, PC y MS office), para llevar a cabo las pruebas.

Con el fin de establecer una base de comparación entre la voz natural y la sintética, se incluyeron aleatoriamente el 25

Los resultados obtenidos en la tasa de inteligibilidad (IR) con los evaluadores tanto para la voz natural como para la sintética se muestran en la tabla 5.

Tipo de voz	IR
Natural	83,45 %
Sintética	88,56 %

Tabla 5: Resultados de la tasa de inteligibilidad.

Mientras que los resultados obtenidos con los mismos evaluadores en la calificación media de opinión (MOS) son los mostrados en la tabla 6.

Tipo de voz	MOS
Natural	3,75 ± 1,2
Sintética	4,01 ± 1,09

Tabla 6: Resultados de la calificación media de opinión (MOS).

Adicionalmente, en la Figura 7 se muestra la compración entre las distribuciones de MOS para la voz natural y sintética.

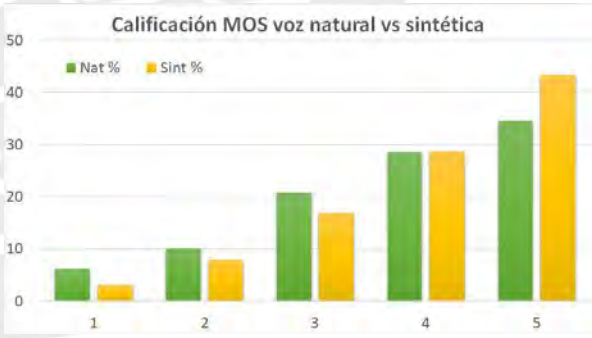

Figura 7: Distribución en porcentaje de MOS para voz natural y sintética.

5.4 Discusión

Finalmente, los resultados fueron positivos. Se esperaba obtener una tasa de inteligibilidad igual o superior al 90 %, y se alcanzó una calificación del 88.56 %. La métrica MOS cumplió con las expectativas, alcanzando una calificación de 4.01.

Sin embargo, al comparar los resultados por su origen, la voz sintética demostró un mejor desempeño que la voz natural tanto en IR como en MOS.

Además de los resultados cuantitativos, se recopilaron observaciones más subjetivas sobre la calidad de la voz sintética. Se señalaron aspectos como la entonación ambigua, el poco respeto a los signos de puntuación y diferencias en la pronuciación.

Desde un enfoque más objetivo para evaluar la voz, durante la síntesis de las frases para las evaluaciones, se generaron las gráficas de alineamiento encoder-decoder correspondientes. Al analizarlas una por una, se identificaron los sonidos que no se generaron adecuadamente con mayor frecuencia y se identificaron las posibles causas. Para el caso de los sonidos *titai, tiai* y *wai* no existen sufcientes frases en el set de datos para un correcto aprendizaje del modelo. Por otro lado, los sonidos *iki, nai, non, tian, ani, ja, ea, baon, kon, xe* y *noa* tienen diferencias en su pronunciación en varios clips de audio del corpus.

6 Conclusiones y trabajos futuros

Se diseñó, desarrolló y evaluó con éxito un modelo de síntesis de texto a voz (TTS) para el Shipibo-Konibo, una lengua indígena peruana de pocos recursos.

Se recopiló un corpus de más de 4 horas de grabaciones y 3025 frases debidamente etiquetadas, listas para ser utilizadas en tareas de procesamiento del lenguaje natural.

Se probó con éxito el método de entrenamiento para los dos componentes del modelo TTS: el predictor de espectrogramas Tacotron 2 y el vocoder HiFi-GAN.

Se logró una tasa de inteligibilidad superior al 88 % y una calificación media de opinión de 4. Estos resultados, considerando la limitación de recursos, son significativos para un modelo TTS llegando a superar la calificación de la voz natural en las mismas pruebas realizadas.

El rendimiento del modelo para esta lengua de la familia Pano abre la posibilidad de emplearlo en futuros desarrollos de modelos TTS para otras lenguas de esta misma familia, así como para otras lenguas amazónicas.

En trabajos futuros, se buscará mejorar la calidad del corpus obtenido por medio de la mejora de las condiciones de registro de audio, incorporando frases con mayor énfasis en puntuación, exclamaciones o preguntas. Además, se prestará especial atención a las frases con los sonidos donde el modelo presentó dificultades.

Referencias

Willem F. H. Adelaar. 1991. The endangered languages problem: South america. In R. H. Robins and E. M. Uhlenbeck, editors, *Endangered Languages*, pages 45–92. New York: Berg.

Koshi Shinanya Ainbo. 2005. El testimonio de una mujer shipiba. *Perú: Editorial e imprenta UNMSM*.

Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. 2017. Deep voice: Real-time neural text-to-speech. In *International conference on machine learning*, pages 195–204. PMLR.

BDPI-MINCUL. 2018. *Ficha de la lengua Shipibo-Konibo*. Ministerio de cultura del Perú, Lima, Perú.

Etta Becker-Donner. 1962. Guaporé-gebiet. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research*, 5:146–150.

John G Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl. 2013. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment. *journal of the audio engineering society*, 61(6):366–384.

Lyle Campbell and Kenneth Rehg. 2018. Introduction. In Lyle Campbell and Kenneth Rehg, editors, *The Oxford Handbook of Endangered Languages*, pages 1–18. Oxford: Oxford University Press.

Arthur Capell. 1962. Linguistic research needed in australia. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research*, 5:23–28.

Rodrigo Cámara-Leret and Jordi Bascompte. 2021. Language extinction triggers the loss of unique medicinal knowledge. *Proceedings of the National Academy of Sciences*, 118:e2103683118.

Ashwini Dasare, KT Deepak, Mahadeva Prasanna, and K Samudra Vijaya. 2022. Text to speech system for lambani-a zero resource, tribal language of india. In *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Ankur Debnath, Shridevi S Patil, Gangotri Nadiger, and Ramakrishnan Angarai Ganesan. 2020. Low-resource end-to-end sanskrit tts using tacotron2, waveglow and transfer learning. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–5. IEEE.

Nicholas Evans and Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–492.

T Gopalakrishnan, Syed Ayaz Imam, and Archit Aggarwal. 2022. Fine tuning and comparing tacotron 2, deep voice 3, and fastspeech 2 tts models in a low resource environment. In *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, pages 1–6. IEEE.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. Glottolog 4.4. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at http://glottolog.org. Accessed on 2021-05-20.

INEI. 2018. *Perú: Resultados definitivos de los Censos Nacionales 2017*. Instituto Nacional de Estadística e Informática, Lima, Perú.

Aleksandr E. Kibrik. 1991. The problem of endangered languages in the ussr. *Diogenes*, 153:67–83.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):1–10.

Michael E. Krauss. 2007. Mass language extinction and documentation: The race against time. In Osahito Miyaoka, Osamu Sakiyama, and Michael Krauss, editors, *Vanishing Languages of the Pacific Rim*, pages 3–24. Oxford: Oxford University Press.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713.

Rene Cedillo Pantoja. 2023. taco2-checkpoints. https://huggingface.co/datasets/rmcpantoja/taco2-checkpoints/tree/main/es,[Accesed: (2024-02-09)].

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Antony W Rix. 2003. Comparison between subjective listening quality and p. 862 pesq score. *Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN'03), Prague, Czech Republic*.

Bonny Sands. 2017. The challenge of documenting africa's least known languages. In Jason Kandybowicz and Harold Torrence, editors, *Africa's Endangered Languages: Documentary and Theoretical Approaches*, pages 11–38. Oxford: Oxford University Press.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

SIL. Accessed March 14 2024. *Ethnologue: Languages of the World*. SIL International, https://www.ethnologue.com/.

Doris Stone. 1962. Urgent tasks of research concerning the cultures and languages of central american indian tribes. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research*, 5:65–69.

Morris Swadesh. 1960. Problems in language salvage for prehistory. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research*, 3:15–19.

Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. 2017. Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv preprint arXiv:1707.06588*.

Pilar M. Valenzuela. 2003. *Transitivity in Shipibo-Konibo Grammar*. Ph.D. thesis, Eugene: University of Oregon.

Stefan Wurm. 1956. Die dringendsten linguistischen aufgaben in neuguinea. In *Ethnologica, seconde partie et rapport général*, volume III of *Actes du IVe congrès international des sciences anthropologiques et ethnologiques, Vienne, 1-8 septembre 1952*, pages 289–292. Wien: Adolf Holzhausens, Wien.

Stephen A. Wurm. 1991. Language death and disappearance: Causes and circumstances. *Diogenes*, 39(153):1–18.

Andrzej Zaborski. 1970. Cushitic languages - an unexplored subcontinent. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research*, 12:119–128.

ANEXOS

I. Fuentes

Estas son las fuentes para el modelo de síntesis de voz (TTS) en lengua shipibo-konibo. Incluye el set de datos, los notebooks para el entrenamiento del predictor de espectrogramas Tacotron 2 y del vocoder HiFi-GAN, así como un notebook de inferencia para la generación de audio sintético.

Set de datos Shipibo-Konibo

https://drive.google.com/file/d/1QpVnGTuXiSFM0T1M4u_mrjDonwKp_2aY/view?usp=sharing

Notebooks de colab del desarrollo del modelo

1. *Entrenamiento*

- Predictor Tacotron 2

  https://colab.research.google.com/drive/1DpEyXxNCo_LyP6YLxYtwaO6poemNM58F?usp=sharing

- Vocoder HiFi-GAN

  https://drive.google.com/file/d/1VymK2GouwIxGklN5gnZnteFYd6rQvYlD/view?usp=sharing

2. *Inferencia*

   https://colab.research.google.com/drive/10aNgyQ47kVsbaNpX1B-4Y1KEdY-upNYM?usp=sharing

II. Protocolo para la Recolección de Datos de Texto y Audio en Lengua Shipibo-Konibo

1. Introducción

El objetivo del presente protocolo es establecer procedimientos estandarizados para una eficiente recolección de datos de audio y texto para la lengua shipibo-konibo con el objetivo de estructurar un corpus suficiente para su uso en tareas de procesamiento del lenguaje natural (PLN) como la síntesis de texto a voz (TTS) o la generación de voz a partir de texto (STT) especialmente enfocado en lenguas de pocos recursos.

2. Criterios de Selección de Participantes

Con base en experiencias previas en la creación de modelos TTS, se establecen los siguientes criterios para la selección de los participantes:

- ***Lengua materna:*** Se seleccionarán hablantes nativos (hombres y mujeres) de preferencia originarios de las comunidades shipibo-konibo de la cuenca del río Ucayali, con el objetivo de minimizar la influencia del español y preservar la autenticidad de la pronunciación y el acento.
- ***Alfabetización:*** Se priorizará a personas alfabetizadas en lengua shipibo-konibo según el estándar oficial establecido por el Ministerio de Educación del Perú.
- ***Experiencia:*** Es preferible que los participantes tengan experiencia en tareas de locución o grabación de audio.
- ***Educación:*** Se recomienda seleccionar participantes con instrucción superior, lo cual facilita el proceso de grabación y seguimiento de instrucciones.
- ***Disponibilidad:*** Los participantes deben contar con disponibilidad para múltiples sesiones de grabación, dada la necesidad de distribuir la carga de trabajo para evitar fatiga.

3. Recolección de Datos Textuales

El corpus textual se compondrá de frases, oraciones y párrafos representativos del uso formal del idioma. Dado que la normalización del alfabeto shipibo-konibo se implementó oficialmente en 2015, el acceso a fuentes escritas es limitado. Las siguientes fuentes han sido identificadas como relevantes:

- ***Cuentos traducidos:*** La traducción de 'El Principito', disponible en el repositorio de la Alianza Francesa (https://www.calameo.com/alianza-francesa-lima/read/006559946768aaaef50da).
- ***Textos educativos:*** Materiales del Ministerio de Educación del Perú para la alfabetización de pueblos originarios (https://repositorio.minedu.gob.pe/handle/20.500.12799/3538) .
- ***Textos religiosos:*** La traducción de la Biblia al shipibo-konibo, aunque con un lenguaje que puede no reflejar el habla cotidiana (https://www.scriptureearth.org/00spa.php?idx=154&language=Shipibo-Konibo&iso_code=shp).

Una vez recopilados los textos estos deben ser separados en frases que contengan de 3 a 16 palabras para su posterior grabación. De experiencias anteriores es recomendable tener especial cuidado en los signos de puntuación como los puntos y comas. Además, se debería incluir la mayor cantidad posible de expresiones exclamativas e interrogativas posibles con el fin de mejorar la calidad del corpus a generar.

4. Recolección de Datos de Audio

El proceso de grabación de audio debe estructurarse de manera que no resulte agotador para los participantes. Se recomienda realizar sesiones de no más de 2 horas, con descansos cada 40 minutos para mantener la consistencia de la entonación y el timbre vocal.

- ***Condiciones de grabación:*** Siempre que sea posible, se debe realizar la grabación en cabinas insonorizadas para minimizar el ruido de fondo.
- ***Duración de los clips:*** Los audios deben tener una duración preferentemente de entre 2 y 10 segundos.
- ***Formato de archivo:*** Los clips de audio deben estar en formato WAV, sin compresión, con una frecuencia de muestreo de al menos 22050 Hz y una profundidad de 16 bits.

5. Estructura del Corpus y Herramientas de Gestión

Texto

Las transcripciones de todas las frases deberán almacenarse en un archivo en formato CSV utilizando la codificación UTF-8, garantizando así la compatibilidad con diversas herramientas de procesamiento de datos. Para la manipulación de este tipo de archivos, se recomiendan aplicaciones como **LibreOffice** (para entornos fuera de línea) o **Google Sheets** (para trabajo colaborativo en la nube). Se sugiere nombrar el archivo como *"transcription.csv"*, siguiendo la estructura de columnas descrita a continuación.

I. ***ID:*** Para mantener un orden se debe usar el nombre del dataset "SK001" seguido del número correlativo del clip de audio al que corresponde. Por ejemplo, el cuarto clip de audio sería "SK001-0004".
II. ***Transcripción:*** Texto hablado en el clip de audio correspondiente.

Basado en la experiencia anterior se recomienda agregar la mayor cantidad posible de frases que contengan los siguientes sonidos**: titai, tiai y wai** ya que mostraron dificultades.

Audio

Cada fila en el archivo *"transcription.csv"* debe estar asociada a un clip de audio correspondiente. El nombre de cada archivo de audio debe seguir la convención: SK001-XXXX.wav, donde XXXX representa el número correlativo del clip de audio.

Se recomienda utilizar grabadoras de audio profesionales como **Zoom** o **Tascam** para la captura del sonido. En cuanto al software de edición, se sugieren opciones de código abierto como **Reaper** o **Audacity** para el procesamiento y ajuste de los clips.

6. Procedimiento de Grabación

Para asegurar la obtención de datos de alta calidad y consistencia, se recomienda seguir el siguiente procedimiento estructurado:

I. ***Preparación de las frases:*** Asegúrese de contar con la lista completa de frases revisadas previamente, evitando así errores y optimizando el tiempo durante la sesión de grabación.
II. ***Configuración de equipos:*** Disponga los equipos de grabación necesarios y posicione el micrófono a una distancia de entre 15 y 30 centímetros del orador. Ajuste el software de grabación según los parámetros recomendados y ofrezca al orador y al operador comodidad suficiente para facilitar el proceso.
III. ***Pruebas iniciales:*** Realice grabaciones de prueba y verifique la calidad del audio, ajustando la ganancia si es necesario.
IV. ***Instrucciones al orador:*** Proporcione directrices claras sobre la forma de hablar, indicando que deben mantener una velocidad natural y una pronunciación clara. Si se comete un error, solicite al orador que haga una pausa y repita la frase desde el inicio.

V. ***Enfoque en sonidos específicos:*** Informe al orador que preste especial atención a la pronunciación de los siguientes sonidos: ***iki, nai, non, tian, ani, ja, ea, bain, kon, xe, noa y xon*** debido a que estas pronunciaciones han mostrado inconsistencias en experiencias previas.

VI. ***Énfasis en la entonación:*** Instruya al orador para que ponga énfasis en la entonación, especialmente en frases exclamativas e interrogativas.

VII. ***Monitoreo en tiempo real:*** Escuche las grabaciones mientras se realizan para identificar problemas como ruido de fondo, interrupciones o pronunciaciones inexactas. Solicite la repetición de cualquier segmento con errores, pausas largas o ruidos inesperados.

VIII. ***Revisión preliminar:*** Escuche nuevamente las grabaciones para verificar que cumplen con los estándares de calidad. Marque los archivos que presenten problemas y evalúe si es necesario volver a grabarlos.

IX. ***Almacenamiento seguro:*** Una vez finalizada la sesión, transfiera los archivos de audio a un almacenamiento seguro, preferentemente con copias tanto en la nube como en medios locales.

X. ***Revisión final:*** Realice una segunda revisión con un equipo especializado para garantizar la calidad de todas las grabaciones. Filtre los audios que no cumplan con los estándares y compile una lista de clips que requieren ser regrabados.

7. **Normalización y Validación**

I. ***Normalización de los audios:*** Tras finalizar las sesiones de grabación, ajuste los clips de audio para estandarizar el volumen y la velocidad, si es necesario, durante el proceso de edición.

II. ***Verificación de correspondencia:*** Asegúrese de que cada archivo de audio en formato WAV esté correctamente asociado con su correspondiente transcripción en el archivo transcription.csv, verificando que los nombres coincidan.

III. ***Organización y empaquetado:*** Almacene todos los clips de audio en una carpeta denominada *"wavs"*. Una vez organizados, comprima la carpeta junto con el archivo *"transcription.csv"* en un archivo ZIP para facilitar su almacenamiento y transferencia.