Revista Causalidad

Artículo https://doi.org/10.70929/caui3.v1i2.0016

Traducción de voz a lenguaje de señas por medio de Aprendizaje Profundo para personas con discapacidad auditiva

Voice-to-sign language translation using deep learning for people with hearing impairments

Recibido: 12, junio 2025

Aceptado: 24, junio 2025

Publicado online: 18, julio 2025

Como Citar: J. R. Alvarado Cadena, "Traducción de voz a lenguaje de señas por medio de Aprendizaje Profundo learning para personas con discapacidad auditiva", *Revista Digital Científica Causalidad (caui3)*, vol. 1, no. 2, pp. 46–56, Jul. 2025, doi: 10.70929/caui3.v1i2.0016.

ISSN-e: 3073-1518

Jorge Alvarado Cadena[1]

Resumen:
Las personas con discapacidad se enfrentan a desafíos históricos; en específico, este trabajo se enfoca en la discapacidad auditiva y la importancia de la lengua de señas. Se propone desarrollar un sistema de traducción de voz a lengua de señas, utilizando tecnologías avanzadas como algoritmos de aprendizaje profundo. El objetivo es alcanzar una mejora en la comunicación y accesibilidad de las personas con discapacidad auditiva, especialmente en entornos donde predomina el lenguaje oral. Destacando la problemática en la comunicación efectiva en Ecuador, tanto en contextos educativos como laborales. Este enfoque abarca la creación y movimiento de un avatar, la conversión de voz a texto por medio de tecnologías Deep Learning, el uso de un diccionario de lengua de señas por sucesión de imágenes. Finalmente, para evaluar el trabajo, se ha ejecutado y evaluado una muestra de 50 personas, con lo que se ha demostrado una alta efectividad en la traducción, destacando la precisión de la inteligencia artificial, la viabilidad en computadoras de bajos recursos y la diversidad de tecnologías aplicables.
Palabras Clave: Lengua de señas, Aprendizaje profundo, Python, MakeHuman.

Abstract:
People with disabilities face historical challenges; specifically, this work focuses on hearing impairment and the importance of sign language. The aim is to develop a voice-to-sign language translation system using advanced technologies such as deep learning algorithms. The goal is to improve communication and accessibility for people with hearing disabilities, especially in environments where spoken language predominates. It highlights the problem of effective communication in Ecuador, both in educational and work contexts. This approach encompasses the creation and movement of an avatar, voice-to-text conversion using deep learning technologies, and the use of a sign language dictionary through a sequence of images. Finally, to evaluate the work, a sample of 50 people was tested and evaluated, demonstrating high effectiveness in translation, highlighting the accuracy of artificial intelligence, the viability on low-resource computers, and the diversity of applicable technologies.
Keywords: Sign language, Deep learning, Python, MakeHuman.

[1]Instituto Superior Tecnológico Carlos Cisneros, Riobamba, Ecuador.
Autor de correspondencia: jorge.alvarado@istcarloscisneros.edu.ec

1 Introducción

A lo largo de la historia, las personas con discapacidad han enfrentado obstáculos estructurales significativos en la exigencia del reconocimiento de sus derechos a la igualdad, la inclusión social y el respeto a su dignidad. Este fenómeno ha derivado recurrentemente en procesos de segregación, tanto en el ámbito familiar como del tejido social general, lo que ha provocado la privación sistemática de oportunidades y el impedimento para el desarrollo de una vida bajo condiciones de equidad y plena participación [1], [2], [3], [4].

En específico, dentro de las diversas las condiciones discapacitantes, la hipoacusia —definida como el déficit parcial o total de la capacidad auditiva, de carácter uni o bilateral, que afecta al sistema sensorial encargado de la transducción y procesamiento de estímulos sonoros— constituye una variable de estudio relevante. En el dominio de la comunicación, las personas con esta condición pueden presentar intercambios dialógicos caracterizados por una estructura lingüística de desarrollo restringido o fragmentado, lo cual genera una barrera comunicativa efectiva que limita su interacción con la sociedad mayoritariamente normo-oyente [5], [6].

La hipoacusia constituye una condición de salud prevalente que afecta a un segmento poblacional significativo en el contexto ecuatoriano [7], [8]. En consecuencia, los individuos que presentan esta condición afrontan barreras significativas en los actos comunicativos de la vida diaria, generadas por una asimetría en las competencias lingüísticas, lo que es particularmente evidente durante las interacciones con pares que no dominan sistemas de comunicación visogestual, como la lengua de señas.

Existe un amplio conjunto de investigaciones enfocado en desarrollar soluciones tecnológicas para esta problemática, así, de acuerdo con la investigación de Carguacundo y Constante [9] donde, se desarrolló una aplicación móvil diseñada para traducir texto y voz a lengua de señas con el objetivo de facilitar la comunicación entre personas sordas y oyentes, así como promover la adquisición de la lengua de señas. En González y Chamorro [10], se documenta el desarrollo e implementación de un aplicativo con funciones de interpretación a lengua de señas, concebido para facilitar la realización de gestiones básicas en un marco de mayor inclusividad; así, la ejecución de este proyecto se estructuró en un proceso secuencial de seis fases. Finalmente, los investigadores Peleo, D. et al. [11] se centraron en el desarrollo de un software embebido para traducir voz a lengua de señas en tiempo real, con el objetivo de salvar la barrera de comunicación entre profesores oyentes y estudiantes con discapacidad auditiva.

En este contexto, el objetivo de este proyecto se enfoca en la optimización de la calidad de vida, tanto para la población con discapacidad auditiva como para la sociedad en general, a través de la facilitación de una comunicación eficaz y fluida en sus contextos pertinentes (esto constituye un recurso de apoyo complementario). Ya sean personas que proveen servicios o de los servidores públicos, el sistema contribuye a la emisión y recepción de mensajes de forma estructurada y asertiva. En este sentido, este trabajo presenta un sistema que pretende resolver una limitación operativa significativa para las personas con hipoacusia; disminuyendo así, la dependencia de un acompañante para gestiones administrativas o actos de consumo. El propósito es crear mejores opciones hacia las personas con discapacidad auditiva, permitiendo que de esta manera no existan barreras de comunicación y se convierta en un problema.

Finalmente, la sección 2 realiza una breve descripción del planteamiento del problema y la a metodología seguida en el trabajo. La sección 3 describe los resultados alcanzados. La sección 4 realiza una breve discusión del trabajo y, finalmente se presentan las conclusiones.

2 Metodología y planteamiento del problema

En esta sección se describe de forma breve los diferentes medios aplicaciones y herramientas específicas que se utilizaron a lo largo del proyecto. De esta forma, el propósito es

crear mejores opciones hacia las personas con discapacidad auditiva, permitiendo que no existan barreras de comunicación y se convierta en un problema.

El objetivo es realizar una breve descripción de como un sistema diseñado para convertir una señal de audio presentada como entrada, se puede convertir en un resultado textual cuyo contenido refleje el mensaje contenido en la entrada de audio. Por lo que se ha acudido a las bondades de la arquitectura secuencia a secuencia [12], [13] muy utilizada en aplicaciones como Traducción automática, Generación de titulares de noticias, Resumen de texto, Conversión de voz a texto, etc. En específico, esta arquitectura cuenta de 2 partes, codificador (Encoder) y decodificador (Decoder).

En la Figura 1 ilustra la arquitectura del modelo, compuesta por dos módulos principales. El primero, denominado Codificador, recibe como entrada una secuencia y genera una representación vectorial continua (a menudo denominada contexto o estado latente) de la misma. El segundo módulo, el Decodificador, utiliza esta representación vectorial como su estado inicial o entrada contextual para generar, de forma autoregresiva, la secuencia de salida en cada paso temporal. El conjunto global del modelo puede ser concebido como una arquitectura secuencia a secuencia, permitiendo que la longitud de la secuencia de entrada pueda variar en comparación con la de la secuencia de salida. Este tipo de modelos ha ganado una importancia destacada en el ámbito del Procesamiento del Lenguaje Natural (PLN) [14], [15], [16].

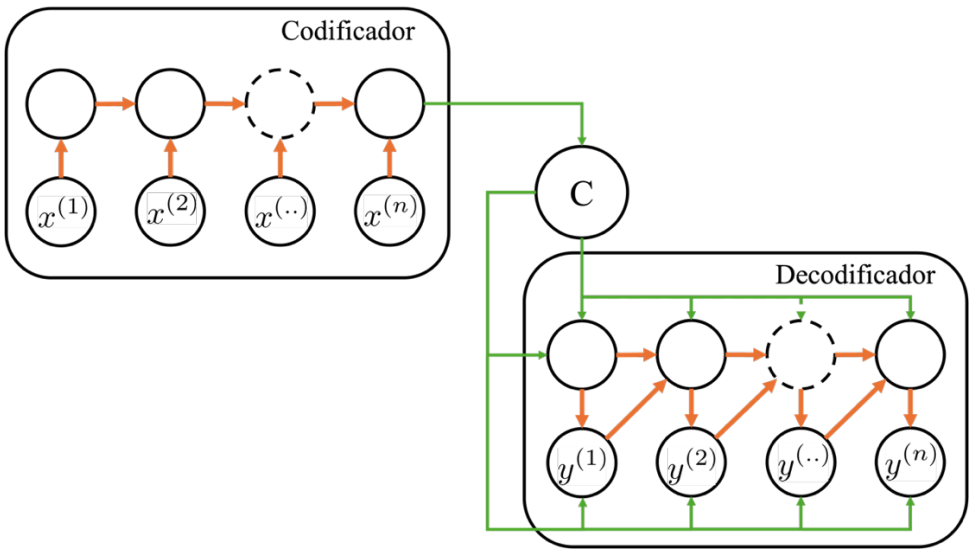


Figura 1. Arquitectura Secuencia a Secuencia.

La Figura 2 esquematiza el proceso de transducción y conversión de señales acústicas utilizadas en el estudio. Al inicio un transductor electroacústico (es este estudio un micrófono) convierte las variaciones de presión sonora (ondas mecánicas) en una señal eléctrica analógica proporcional, traducido como voltaje variable en el tiempo. Posterior, para su procesamiento digital, esta señal analógica es sometida a un proceso de conversión analógico-digital (ADC), mediante el cual se muestrea, cuantifica y codifica en un formato binario, generando así una señal de audio digital.

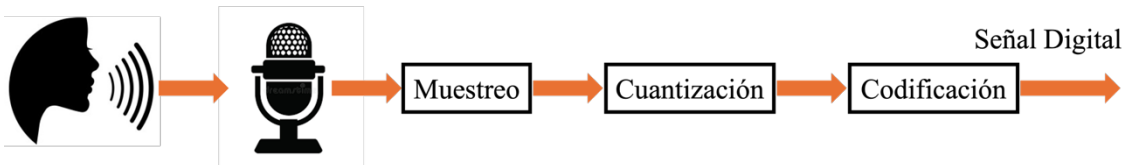


Figura 2. Transformación de señal sonora a señal digital.

En la Figura 3 describe la configuración de interacción usuario-plataforma y el proceso de conversión de voz a texto. Al iniciar una grabación, la señal de audio resultante se almacena parametrizada por su frecuencia de muestreo y configuración de canales. La frecuencia de muestreo mínima se establece en 16000 Hz, conforme a las especificaciones del servicio. Cabe señalar que, según la Organización Internacional de Normalización (ISO), el rango de frecuencias audibles para el oído humano se sitúa entre 125 Hz y 20000 Hz. Respecto a la configuración de canales, se asigna el valor 2 para representar audio estereofónico. Finalmente, la señal se codifica y almacena en formato de archivo WAV.

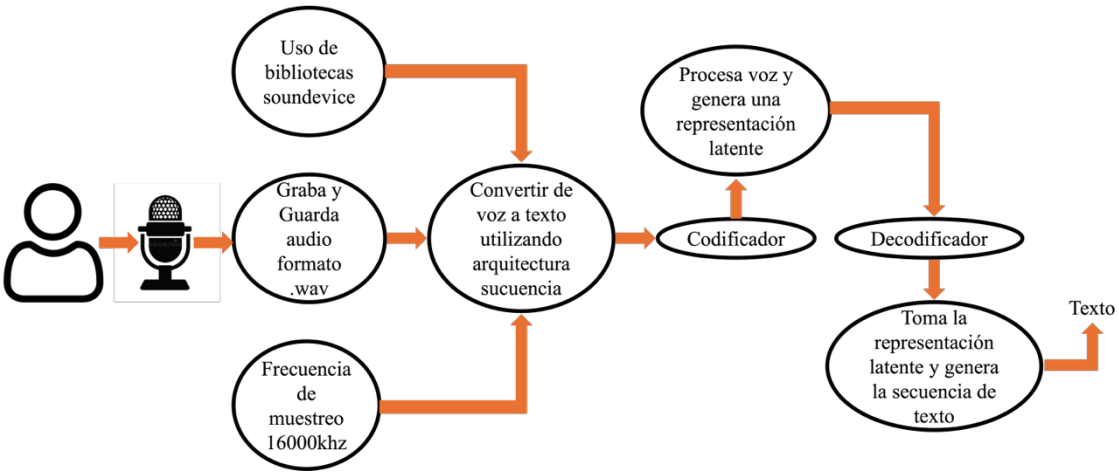


Figura 3. Diagrama de caso de uso de la conversión de voz a texto.

A continuación, se procede la conversión del audio por medio de redes neuronales recurrentes donde se utiliza la arquitectura secuencia a secuencia.

Por una parte, el codificador utiliza un conjunto de datos de entrenamiento que contienen audios previamente escritos (frases definidas en el estudio) para después aprender las representaciones temporales, también debe contener una variedad de estilos de habla, acentos y condiciones acústicas con respecto a cada palabra y así realizar reconocimiento patrones. Entonces, durante el entrenamiento se utilizan algoritmos de optimización (descenso de gradiente estocástico) que son parámetros que se ajustan progresivamente para disminuir la diferencia entre la salida predicha y la salida real, como resultado se obtiene un valor de salida que es utilizado como entrada para la red neuronal del decodificador.

El decodificador es entrenado utilizando como entrada la salida contextual generada por el codificador, el cual es emparejado con su transcripción textual de referencia para el cálculo de la función de pérdida. La arquitectura neuronal del decodificador está específicamente diseñada para generar secuencias de texto lingüísticamente coherentes y semánticamente precisas. Durante este proceso, sus parámetros internos son optimizados mediante algoritmos de retropropagación del error, con el objetivo de minimizar la discrepancia entre las secuencias de texto predichas por el modelo y las transcripciones reales del conjunto de entrenamiento.

En la Figura 4 se muestra el proceso de búsqueda de la imagen según el texto generado para luego ejecutarlo y alcance una visualización del usuario en la lengua de señas mediante una interfaz gráfica.

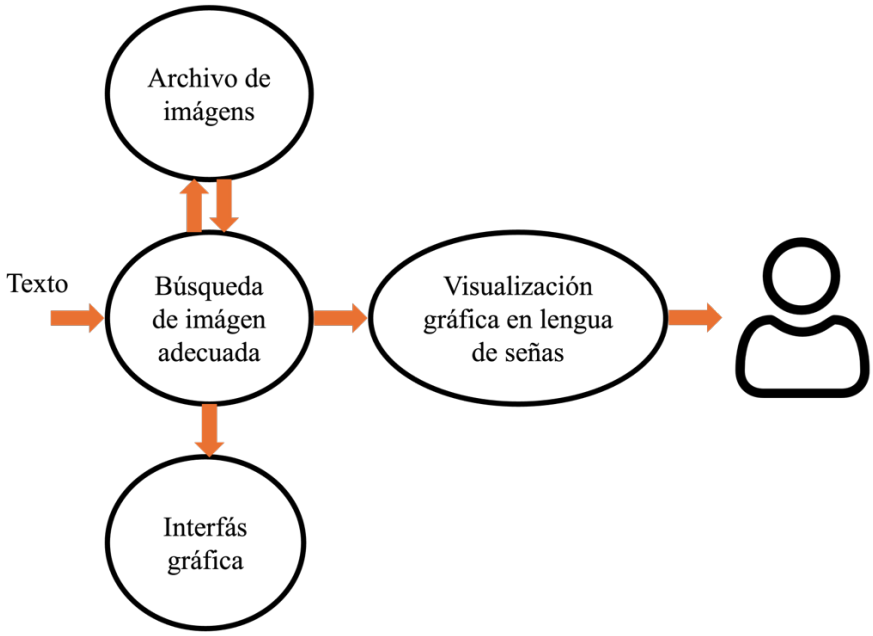
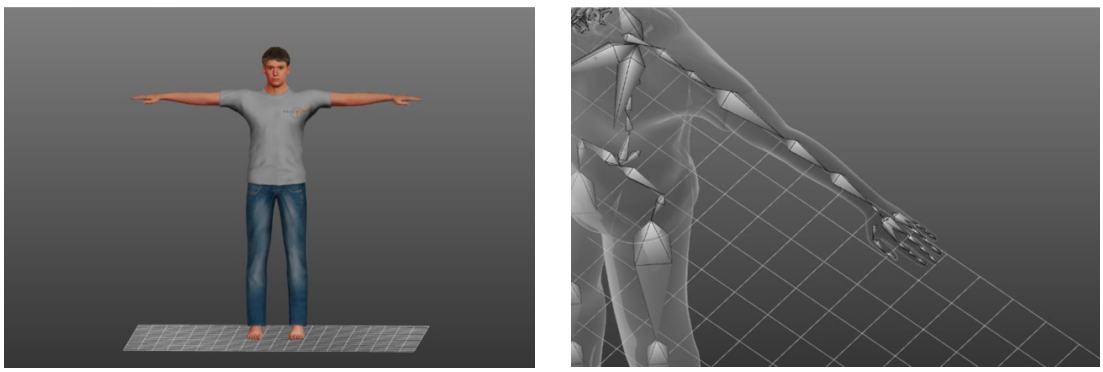


Figura 4. Diagrama de caso de uso de conversión de texto a lengua de señas.

2.1 Diseño del Avatar

En específico, se ha creado un avatar que interactúe entre el software y el usuario. En la Figura 5a se apriencia el avatar desaroollado en MakeHuman. Mienmtras, la Figura 5b visualiza el esqueleto que se le agrega al avatar para darle movimiento, es importante tener en cuenta que como el tema del proyecto es la lengua de señas, se enfoca el esqueleto, en específico en los brazos y manos, ya que es la principal fuente de comunicación de las personas con discapacidad auditiva.



a) b)

Figura 5. Avatar del Software. a) Avatar personalizado en MakeHuman. b) Esqueleto del avatar para el movimiento de lengua de señas.

Para realizar el movimiento del avatar, se exporta el archivo del software MakeHuman para luego importarlo al aplicativo Blender, que tiene como finalidad soportar dispositivos informáticos de bajos recursos. Así, Los movimientos de las manos se realizan de manera directa, se mueve cada punto de articulación de acuerdo con su eje y su rotación según la necesidad, como se ve en la Figura 6.

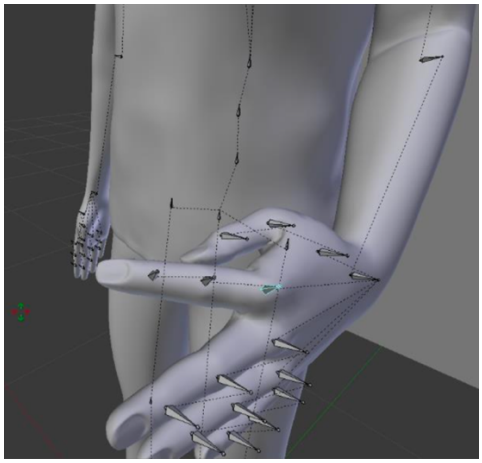


Figura 6. Movimiento de las articulaciones del avatar.

2.2 Diccionario de lengua de señas

Es importante destacar que las lenguas de señas que se realizaron han sido tomadas como referencia del Consejo Nacional de Igualdad de Discapacidades (CONADIS) [17], quienes crearon el diccionario de lengua de señas a través del ecuatoriano Gabriel Román y fueron desarrollados con el apoyo de la Federación Nacional de Sordos del Ecuador (FENASEC) y la Universidad Tecnológica Indoamerica – UTI.

Entonces, en la Figura 7 se observa las capturas de pantalla que se realizan por cada movimiento para luego ser unidas y así formar una lengua de señas.

Figura 7. Imágenes para la creación de una lengua de señas.

3 Resultados

Partiendo de una población total de 2.746.403 habitantes en el cantón Guayaquil, según el último censo del Instituto Nacional de Estadística y Censos (INEC), y ante la impracticabilidad de acceder a la totalidad de la población objeto de estudio debido a limitaciones logísticas y de recursos, se implementó una metodología de muestreo no probabilístico por conveniencia.

Así, la muestra de estudio consiste en 150 participantes seleccionados de manera aleatoria que desconozcan la lengua de señas y no sepan como comunicarse con las personas con discapacidad auditiva. Es importante señalar que, aunque la muestra no representa la población completa de manera exhaustiva, se espera que proporcione información valiosa sobre la usabilidad y eficacia del software desarrollado. Esta estrategia de muestreo se ha seleccionado para maximizar la eficiencia en la recopilación de datos y garantizar la participación de individuos que puedan proporcionar una perspectiva significativa sobre el uso del software en cuestión, así, se ha definido un intervalo de confianza descritos en las ecuaciones (1) y (2).

$$p=\frac{na}{n} \tag{1}$$

$$\text{SE}=\sqrt{\frac{p(1-p)}{n}} \tag{2}$$

$$\textit{Intervalo de Confianza} = p \pm (\textit{valor crítico} \times \text{SE}) \tag{3}$$

donde:

p: proporción estimada
n: tamaño de la muestra
SE: error estándar

Para un nivel de confianza del 90% se utiliza un valor crítico de 1.645 según la tabla de distribución normal estándar, se realiza un ejemplo de cálculo con la frase "Deme su número" a partir de la ecuación (3), mostrando así los valores y resultado de la ecuación según (4) y (5).

$$\textit{Intervalo de confianza} = 0.96 \pm \sqrt{\frac{0.96 \times (1-0.96)}{50}} \tag{4}$$

$$\textit{Intervalo de confianza} = [0.876 \, , \, 1.044] \tag{5}$$

Se toma como referencia el valor mínimo del intervalo de confianza, correspondiente al 87.6%, el cual representa el nivel de confianza mínimo estimado para el rendimiento de reconocimiento de dicha frase. En el Cuadro 1 evidencia que la mayoría de las frases mantienen una tasa mínima de acierto del 95.5%, indicando una precisión elevada. Un subconjunto de casos presenta tasas entre el 85% y el 88%; esta reducción en la precisión se atribuye a errores de transcripción, originados por una vocalización deficiente o por la omisión de énfasis en sílabas tónicas. Finalmente, se registra un caso de efectividad mínima (64.2%) donde la causa principal es la vocalización errónea de la palabra "cédula", siendo transcrita como "célula" o "según". La mitigación de este error implica una articulación más pausada y precisa.

Cuadro 1. Efectividad del rendimiento de cada frase.

Frases	**Número de aciertos**	**Porcentaje de acierto mínimo %**
¿Cuánto dinero necesita?	50	95.5%
Deme su número	47	85.7%
Gracias	50	95.5%
Hola	50	95.5%
Llene estos formularios	50	95.5%
Necesito un préstamo	48	87.6%
No hay sistema	50	95.5%
¿Qué va a pagar?	48	87.6%
Cédula de identidad	37	64.2%
Un momento por favor	50	95.5%

En la Figura 8 se presenta el nivel mínimo porcentual del rendimiento de cada una de las frases y se aprecia un alto rendimiento de factibilidad.

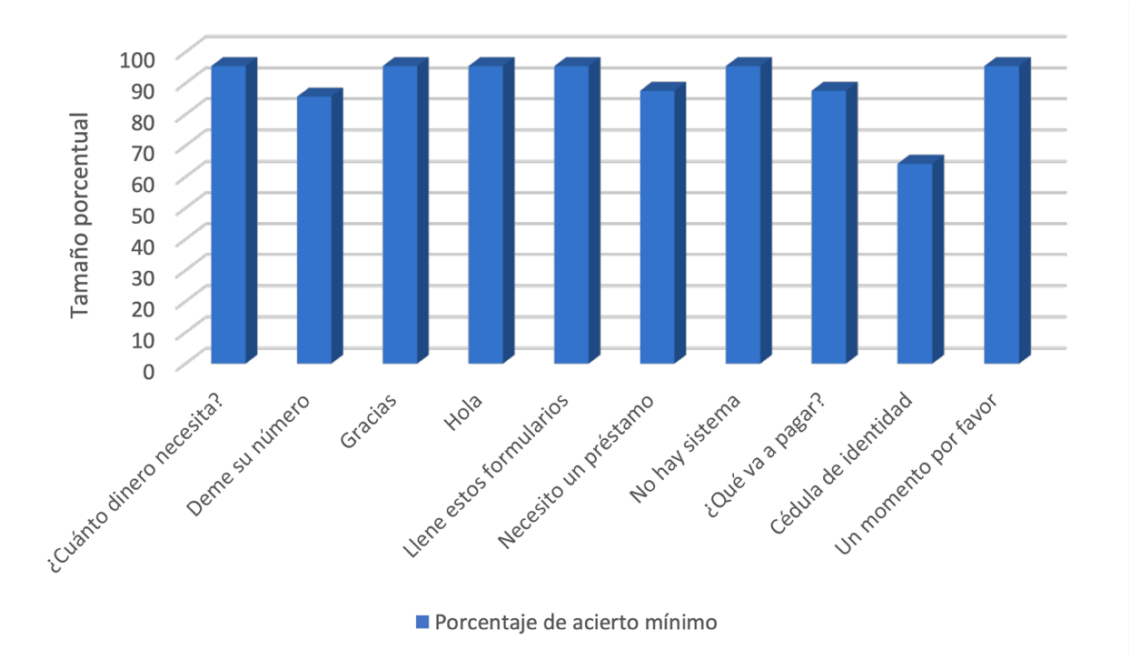


Figura 8. Gráfico porcentual de acierto mínimo en cada frase.

Por otra parte, es importante destacar que, con el objetivo de comparar el rendimiento computacional en tiempo de ejecución, se ha probado las capacidades de 2 computadores de características diferentes. Entonces, se realizó una la medición que inicia desde la finalización de la grabación hasta la aparición de la primera imagen como se muestra en la Tabla 4.

En la Figura 9 presenta un análisis comparativo de tiempos de ejecución mediante un diagrama de barras. Los resultados indican que, si bien el tiempo de respuesta en un dispositivo de recursos limitados no resulta deficiente, se observa una diferencia cuantificable —aunque leve— en el tiempo de procesamiento cuando se compara con el rendimiento obtenido en un dispositivo de recursos óptimos.

Cuadro 2. Comparación del tiempo de ejecución en diferentes dispositivos.

Frases	DESKTOP-LE0CNI Intel(R) Core(TM) Duo PC P750 @ 2.13GHz, 4.00 GB RAM Tiempo de respuesta [s]	Didi Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.71GHz, 6.00GB RAM Tiempo de respuesta [s]
¿Cuánto dinero necesita?	11.02	6.7
Deme su número	7.78	5.8
Gracias	8.6	6.4
Hola	7.3	5.6
Llene estos formularios	7.7	5.6
Necesito un préstamo	6.1	5.6
No hay sistema	9.5	5.9
¿Qué va a pagar?	5.82	5.6

Cédula de identidad	7.2	5.6
Un momento por favor	5.7	5.9

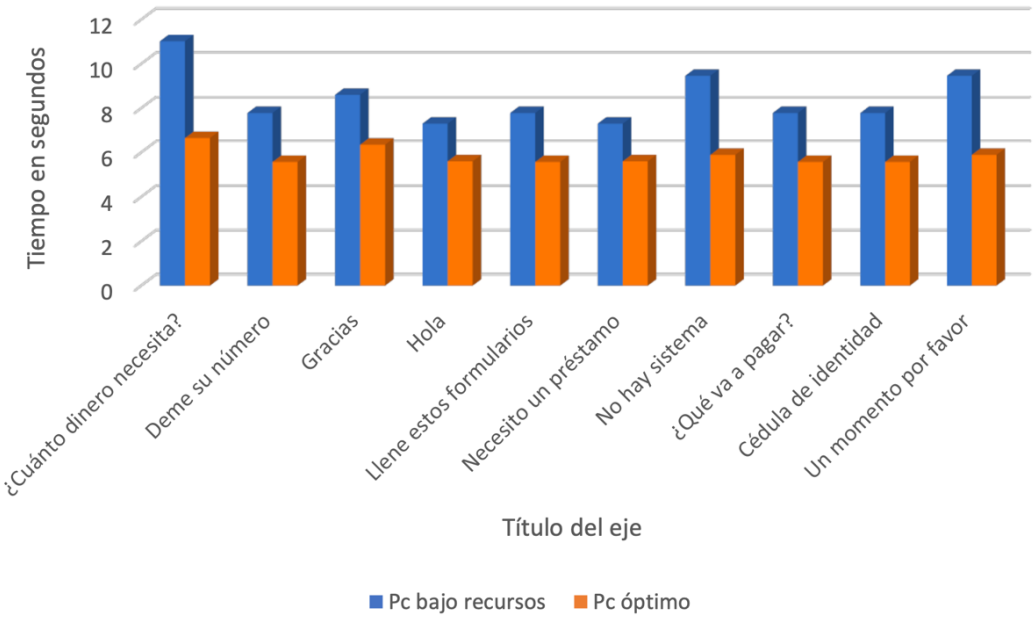


Figura 9. Gráfico comparativo con respecto al tiempo de ejecución en segundos.

Lo que muestra que la aplicación no consume recursos en exceso, por tanto, es accesible en hardware actual y no actual.

4 Conclusiones

El servicio de inteligencia artificial muestra un alto rendimiento en términos de precisión para la tarea de conversión de voz a texto, a pesar de presentar errores residuales atribuibles principalmente a una articulación deficiente o a una tasa de elocución elevada.

Se evidencia que el sistema propuesto es viable para su despliegue en equipos con recursos computacionales limitados, eliminando así una barrera de hardware y facilitando su implementación en contextos donde las personas con discapacidad auditiva requieren un servicio de comunicación eficaz.

Existe un amplio espectro de tecnologías basadas en aprendizaje profundo (Deep Learning) que pueden ser implementadas para abordar necesidades específicas, permitiendo su adaptación a múltiples ámbitos de aplicación y su evolución frente a nuevas demandas emergentes en el tejido social.

Declaración de conflicto de intereses

El autor no ha declarado ningún posible conflicto de intereses en relación con esta investigación, la autoría y/o la publicación de este artículo.

Referencias

[1] A. Franco-Segovia, "La lengua de señas ecuatoriana para la inclusión de los estudiantes con discapacidad auditiva," *Polo del Conocimiento*, vol. 8, no. 2, pp. 458–469, Dec. 2023, doi: 10.23857/pc.v8i2.

[2] J. C. García, "LA DISCAPACIDAD AUDITIVA. PRINCIPALES MODELOS Y AYUDAS TÉCNICAS PARA LA INTERVENCIÓN.," *Revista Internacional de Apoyo a la Inclusión, Logopedia, Sociedad y Multiculturalidad*, pp. 24–36, Dec. 2015, [Online]. Available: http://riai.jimdo.com/

[3] E. M. Ferrero, "Legal language as a barrier to access to justicefor people with disabilities," *Perspectivas*, vol. 11, no. 2, pp. 3–16, Jul. 2021, doi: 10.19137/perspectivas-2021-v11n2a01.

[4] G. M. Campos, "La discapacidad en el lenguaje o el lenguaje para la discapacidad," *Diá-logos*, pp. 78–89, Dec. 2021, doi: 10.5377/dialogos.v23i1.14756.

[5] M. De La Hoz Vásquez, "Barreras en la comprensión lectora en estudiantes con discapacidad auditiva en Educación Superior," vol. 33, no. 26, pp. 77–97, Apr. 2025.

[6] D. Marín, "Desarrollo de la competencia lectora en alumnado con discapacidad auditiva: una revisión bibliográfica," *ReiDoCrea: Revista electrónica de investigación Docencia Creativa*, vol. 8, pp. 142–153, Oct. 2019, doi: 10.30827/Digibug.57749.

[7] J. D. Guerrero-Balaguera and W. J. Pérez-Holguín, "Sistema traductor de la lengua de señas Colombiana a texto basado en FPGA," *DYNA (Colombia)*, vol. 82, no. 189, pp. 172–181, 2015, doi: 10.15446/dyna.v82n189.43075.

[8] D. Quinta Lipe and P. J. Guerrero, "Traducción inteligente de lenguaje de señas mediante aprendizaje automático y visión artificial," *Revista Criterio*, vol. 5, no. 8, pp. 22–43, Apr. 2025, doi: 10.62319/criterio.v.5i8.33.

[9] M. Carguacundo and P. Constante, "Traductor de texto y voz a lengua de señas ecuatoriana a través de un avatar implementado para dispositivos Android," *Infociencia*, vol. 12, no. 1, pp. 20–25, 2018.

[10] E. E. P. Gonzalez and C. R. I. Chamorro, "Desarrollo e implementación de un sistema que permita la accesibilidad de los servicios a personas con discapacidad auditiva/sordos (PcDa/Sordos) por medio de un aplicativo intérprete a lengua de señas en tiempo real," in *Anais do XVIII Congresso Latino-Americano de Software Livre e Tecnologias Abertas (Latinoware 2021)*, Sociedade Brasileira de Computação - SBC, Oct. 2021, pp. 154–157. doi: 10.5753/latinoware.2021.19924.

[11] D. L. C. Peleo, M. M. Patalay, J. M. Dela Cruz, and E. A. Ramirez, "Sign the Voice: An Embedded Software for Voice to Sign Language Translation," in *2024 2nd International Conference on Computing and Data Analytics (ICCDA)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/ICCDA64887.2024.10867405.

[12] I. M. Greca, J. Ortiz-Revilla, and I. Arriassecq, "Diseño y evaluación de una secuencia de enseñanza-aprendizaje STEAM para Educación Primaria," *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*, vol. 18, no. 1, pp. 1–20, 2021, doi: 10.25267/Rev_Eureka_ensen_divulg_cienc.2021.v18.i1.1802.

[13] Á. López-Mota and G. Moreno-Arcuri, "SUSTENTACIÓN TEÓRICA Y DESCRIPCIÓN METODOLÓGICA DEL PROCESO DE OBTENCIÓN DE CRITERIOS DE DISEÑO Y VALIDACIÓN PARA SECUENCIAS DIDÁCTICAS BASADAS EN MODELOS: EL CASO DEL FENÓMENO DE FERMENTACIÓN.," *Revista Bio-grafía Escritos sobre la biología y su enseñanza*, vol. 7, no. 13, p. 109, Dec. 2014, doi: 10.17227/20271034.vol.7num.13bio-grafia109.126.

[14] C. Arana, "Redes neuronales recurrentes: Análisis de los modelos especializados en datos secuenciales," May 2021. [Online]. Available: www.cema.edu.ar/publicaciones/doc_trabajo.html

[15] H. Lyu, N. Sha, S. Qin, M. Yan, Y. Xie, and R. Wang, "Manifold Denoising by Nonlinear Robust Principal Component Analysis."

[16] M. van Gerven and S. Bohte, “Editorial: Artificial Neural Networks as Models of Neural Information Processing,” *Front Comput Neurosci*, vol. 11, Dec. 2017, doi: 10.3389/fncom.2017.00114.
[17] Consejo Nacional de Igualdad de Discapacidades (CONADIS), “Diccionario de Lengua de Señas Ecuatoriana Gabriel Román,” http://www.plataformaconadis.gob.ec/~platafor/diccionario/.