MODELO DE LENGUAJE NATURAL PARA LA TRANSCRIPCIÓN ANONIMIZADA DE GRABACIONES DE AUDIO DEL ESPAÑOL DE COLOMBIA

ANDREA JULIANA PARRA ARIZA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

MODELO DE LENGUAJE NATURAL PARA LA TRANSCRIPCIÓN ANONIMIZADA DE GRABACIONES DE AUDIO DEL ESPAÑOL DE COLOMBIA

ANDREA JULIANA PARRA ARIZA

Trabajo de Grado para optar al título de
Ingeniero de Sistemas

Director:
PhD. Hoover Fabián Rueda Chacón
PhD. en Ingeniería Eléctrica y Computación

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

DEDICATORIA

A toda mi familia: A mis padres, por su apoyo y amor incondicional; a mi hermana, quien siempre ha sido una fuente de aspiración; a Nikka, compañera de mis noches en vela; y mis amigos, cuya presencia ha sido tan firme que ya clasifican como familia.

A todos ustedes, que me han apoyado incluso en los gestos más pequeños; a veces me siento sola como las piedras, pero recuerdo que ellas también son besadas por el mar, y ustedes son mi mar.

AGRADECIMIENTOS

Mi mayor temor antes de finalizar este trabajo de grado era escribir los agradecimientos. No porque no tuviera a quién agradecer, todo lo contrario, son tantas las personas que aparecen en mi mente que sus nombres se entremezclan, y nunca me he sentido tan querida como en este momento. Aún así, intentaré mencionarlos.

Agradezco a mi familia —Alejandra, Claudia, Javier y Nikka— por su amor y por las enseñanzas que me han formado en la persona que soy.

A mis amigos —Antonia, Mariana, Henao, Daniza, Zazu, Almeida, Sebastián, Sofía, Paula, Guarín, Nicolás y muchos más— que con su apoyo constante y sus risas me han hecho sentir acompañada tanto en la tormenta como en la calma.

Al semillero Hands-On Computer Vision, y en especial a su director, Hoover, quien inconscientemente reavivó mi amor por el área y encendió mi cariño por la investigación. Valoro mucho su paciencia y orientación a lo largo de este viaje académico.

Y a todos los que no pude mencionar: también están en mi corazón. Incluso aquellos con quienes crucé apenas unas palabras; por mínimas que hayan sido, las agradezco profundamente.

CONTENIDO

LISTA DE FIGURAS

pág.

LISTA DE CUADROS

pág.

RESUMEN

TÍTULO: MODELO DE LENGUAJE NATURAL PARA LA TRANSCRIPCIÓN ANONIMIZADA DE GRABACIONES DE AUDIO DEL ESPAÑOL DE COLOMBIA[*]

AUTOR: ANDREA JULIANA PARRA ARIZA[**]

PALABRAS CLAVE: *Transformers*, Reconocimiento de Entidades Nombradas, Anonimización, Reconocimiento Automático del Habla, Transcripción.

DESCRIPCIÓN:

El habla, una de las habilidades humanas más esenciales, ha motivado el desarrollo de sistemas de Reconocimiento Automático del Habla (ASR, del inglés *Automatic Speech Recognition*) capaces de convertir el habla en texto escrito. Desde los primeros sistemas de los años 50 hasta modelos modernos basados en Redes Neuronales Profundas, como *Whisper*, los avances han permitido transcripciones multilingües y precisas, así como la integración de tareas de Procesamiento de Lenguaje Natural (NLP) como Reconocimiento de Entidades (NER, del inglés *Named Entity Recognition*) y diarización de hablantes. Sin embargo, estos modelos requieren grandes volúmenes de datos, lo que limita su desempeño en idiomas o variantes con recursos limitados, como el español de Colombia, que presenta acentos y regionalismos poco representados en conjuntos de entrenamiento. Así mismo, el uso de datos reales suele incluir información sensible, como nombres o identificaciones, que dificulta la recopilación e intercambio de estos corpus. En este sentido, contar con un sistema ASR que incorpore mecanismos de anonimización permitiría proteger la privacidad de los hablantes y facilitaría la recolección y distribución de conjuntos de datos. Este trabajo propone desarrollar un modelo de transcripción anonimizada para el español colombiano, incorporando tareas de NLP y marcas de tiempo, con el objetivo de cerrar la brecha entre los modelos existentes y este dialecto, garantizando un desempeño robusto incluso en entornos con datos escasos, con modelos que alcanzan un $7,60\,\%$ de error de transcripción a nivel de palabra (*Word Error Rate*), un F1-score de $60{,}81\,\%$ para NER exacto y un F1-score de $76{,}10\,\%$ en censura.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: PhD. Hoover Fabián Rueda Chacón.

ABSTRACT

TITLE: NATURAL LANGUAGE MODEL FOR ANONYMIZED TRANSCRIPTION OF AUDIO RECORDINGS FROM COLOMBIAN SPANISH*

AUTHOR: ANDREA JULIANA PARRA ARIZA**

KEYWORDS: Transformers, Named Entity Recognition, Anonymize, Automatic Speech Recognition, Transcription

DESCRIPTION:

Speech, one of the most essential human abilities, has driven the development of Automatic Speech Recognition (ASR) systems capable of converting spoken language into written text. From the first systems in the 1950s to modern models based on Deep Neural Networks like *Whisper*, these advances have enabled accurate multilingual transcriptions as well as the integration of Natural Language Processing (NLP) tasks like Named Entity Recognition (NER) and speaker diarization. However, these models require large amounts of data, which limits their performance in languages or variants with scarce resources, such as the Colombian Spanish, which exhibits accents and regionalisms underrepresented in training datasets. Likewise, the use of real-world data often includes sensitive information, such as names or IDs, which makes the collection and sharing of these corpora difficult. In this regard, having an ASR that incorporates anonymization mechanisms will protect the privacy of speakers and facilitate the collection and distribution of data sets. This work proposes the development of an anonymized transcription model for Colombian Spanish, incorporating NLP tasks and time-stamped annotations, aiming to bridge the gap between existing models and this dialect, ensuring robust performance even in low-resource settings, with models achieving a $7{,}60\,\%$ transcription word error rate (WER), an F1-score of $60{,}81\,\%$ for exact NER and an F1-score of $76{,}10\,\%$ for censoring.

---

* Bachelor Thesis

** Faculty of Physical-Mechanical Engineering. School of Computer Science. Advisor: PhD. Hoover Fabián Rueda Chacón.

INTRODUCCIÓN

El habla, una de nuestras habilidades naturales más esenciales, ha sido objeto de estudio por el deseo de construir modelos computacionales capaces de emular las capacidades de comunicación verbal humana. En 1952 con Audrey,[1] se vio el primer sistema para Reconocimiento de Voz Automático (ASR, del inglés *Automatic Speech Recognition*), capaz de distinguir dígitos hablados con más del 90 % de exactitud. Sin embargo, su uso era limitado: con hablantes distintos a su creador, el rendimiento disminuía considerablemente, y el requerir una sala llena de circuitos especializados para cada dígito lo hacía impráctico.

Gracias a los numerosos avances tecnológicos que se han tenido con los años, se han propuesto sistemas complejos de ASR implementados para automatizar distintas tareas: servicios con manejo automático de llamadas como el *Voice Recognition Call Processing* de AT&T,[2] que permite a los usuarios interactuar con sistemas telefónicos mediante reconocimiento de voz, o servicios con capacidades de consulta y asistentes personales como Copilot y Siri.[3] Una de las razones detrás de este rápido avance en el reconocimiento automático del habla es la llegada de modelos basados en Redes Neuronales Profundas (DNN, por sus siglas en inglés *Deep Neural Networks*); en Switchboard, un conjunto de evaluación con conversaciones telefónicas en inglés, varios modelos han alcanzado una exactitud al transcribir similar o mejor que la

---

[1] Katia Moskvitch. *The machines that learned to listen*. https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen. Último acceso: 16 de noviembre de 2025. 2017.

[2] Engineering National Academies of Sciences y Medicine. *Voice Communication Between Humans and Machines*. Applications of Voice-Processing Technology in Telecommunications. Washington, DC: The National Academies Press, 1994. DOI: 10.17226/2308.

[3] Sadeen Alharbi et al. “Automatic Speech Recognition: Systematic Literature Review”. En: *IEEE Access* PP (sep. de 2021), págs. 1-1. DOI: 10.1109/ACCESS.2021.3112535.

humana.[4] Destaca especialmente *Whisper*, un modelo de transcripción multilingüe desarrollado por OpenAI, capaz de generalizar a múltiples *benchmarks* estándar y publicado como código abierto para servir como base en futuras investigaciones sobre el procesamiento robusto del habla.[5] Defínase transcripción en el contexto de ASR, como el proceso de convertir una oración hablada en una secuencia escrita de letras y palabras.[6] Más allá de la transcripción, también se han comenzado a integrar tecnologías basadas en DNN en tareas de Procesamiento del Lenguaje Natural (NLP, del inglés *Natural Language Processing*), tal como el reconocimiento de entidades (NER, del inglés *Named Entity Recognition*), análisis de sentimientos, reconocimiento del lenguaje conversacional, entre otros.[7]

No obstante, para llegar a un buen rendimiento, estos modelos necesitan grandes cantidades de datos, pues un conjunto de datos incompleto, sesgado, o en formatos inconsistentes, afecta el desempeño del modelo al predecir o generar texto.[8] Es por esto que la mayoría de estos avances se han limitado a los idiomas de mayor población hablante. En consecuencia, los idiomas con poca cantidad de habla transcrita o

---

[4] Hemant Yadav y Sunayana Sitaram. "A Survey of Multilingual Models for Automatic Speech Recognition". En: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, jun. de 2022, págs. 5071-5079.

[5] Alec Radford et al. "Robust Speech Recognition via Large-Scale Weak Supervision". En: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul de 2023, págs. 28492-28518.

[6] Harsh Ahlawat, Naveen Aggarwal y Deepti Gupta. "Automatic Speech Recognition: A survey of deep learning techniques and approaches". En: *International Journal of Cognitive Computing in Engineering* 6 (2025), págs. 201-237. DOI: https://doi.org/10.1016/j.ijcce.2024.12.007.

[7] Libo Qin et al. *Large Language Models Meet NLP: A Survey*. 2024.

[8] Imad Zeroual y Abdelhak Lakhouaja. "Data science in light of natural language processing: An overview". En: *Procedia Computer Science* 127 (2018). Proceedings of the first international conference on Intelligent Computing in Data Sciences, ICDS2017, págs. 82-91. DOI: https://doi.org/10.1016/j.procs.2018.01.101.

no transcrita, es decir, los idiomas con recursos limitados, se han quedado atrás.[9] En el caso de Colombia, aunque es un país hispanohablante, posee acentos y regionalismos que los modelos de lenguaje actuales no interpretan con precisión, debido a la escasez de datos transcritos de su variante lingüística. Asimismo, políticas de privacidad dificultan la creación de estos datos, ya que la difusión de esta información puede comprometer identidades sensibles mencionadas, lo que hace necesaria la anonimización de nombres, organizaciones y datos personales.

Con este trabajo de investigación se busca desarrollar un modelo de lenguaje natural para la transcripción anonimizada de grabaciones de audio del español de Colombia, con el objetivo de ayudar a cerrar esta brecha entre los modelos de español actuales y el español de Colombia. En particular, se proyecta incorporar tareas de NLP en la arquitectura del modelo, incluyendo NER, diarización de hablantes y anotación con marcas de tiempo por intervención; manteniendo la anonimidad de los interlocutores.

---

[9] Yadav y Sitaram, ver n. 4.

1. OBJETIVOS

Objetivo general Desarrollar un modelo de lenguaje natural para la transcripción anonimizada, identificación de entidades y anotación con marcas de tiempo de intervenciones en grabaciones de audio del español de Colombia.

Objetivos específicos

1. Recopilar un conjunto de grabaciones de audio del español de Colombia con variedad de hablantes, acentos y expresiones.
2. Diseñar un flujo de trabajo para la transcripción anonimizada, identificación de entidades y anotación con marcas de tiempo de intervenciones en grabaciones de audio del español de Colombia.
3. Implementar un modelo de lenguaje natural por medio de *fine-tuning* supervisado correspondiente al flujo de trabajo diseñado.
4. Evaluar el modelo implementado, calculando métricas de error sobre un subconjunto de grabaciones del español de Colombia con etiquetas de entidades y transcripciones; y midiendo la efectividad de la anonimización en cuanto a la preservación de la información útil y la eliminación de datos sensibles.

2. MARCO DE REFERENCIA

2.1. PROCESAMIENTO DEL LENGUAGE NATURAL (NLP)

El procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés *Natural Language Processing*) abarca el análisis de datos lingüísticos, generalmente en formas textuales, como documentos o publicaciones, mediante métodos computacionales; su objetivo es transformar el lenguaje natural no estructurado en una representación estructurada que pueda ser entendida, analizada o utilizada por una máquina.[10] Dentro de las posibles tareas que abarca NLP se encuentra el entendimiento del lenguaje natural, razonamiento matemático, traducción automática, generación de lenguaje natural, extracción de información estructural (como extracción de entidades), entre otras.[11]

Generalmente, a la hora de trabajar con reconocimiento del habla se busca combinar los resultados generados por el modelo con tareas de NLP para estructurar, interpretar y enriquecer el texto resultante.

2.2. RECONOCIMIENTO DE ENTIDADES NOMBRADAS (NER)

El Reconocimiento de Entidades Nombradas (NER, dado las siglas en inglés *Named Entity Recognition*) es un subcampo de la informática y el NLP que se centra en identificar y clasificar entidades presentes en textos no estructurados, en categorías predefinidas como personas, ubicaciones geográficas y organizaciones. Con el

---

[10] Karin Verspoor y Kevin Bretonnel Cohen. “Natural Language Processing”. En: *Encyclopedia of Systems Biology*. Ed. por Werner Dubitzky et al. New York, NY: Springer New York, 2013, págs. 1495-1498. DOI: 10.1007/978-1-4419-9863-7_158.

[11] Qin et al., ver n. 7.

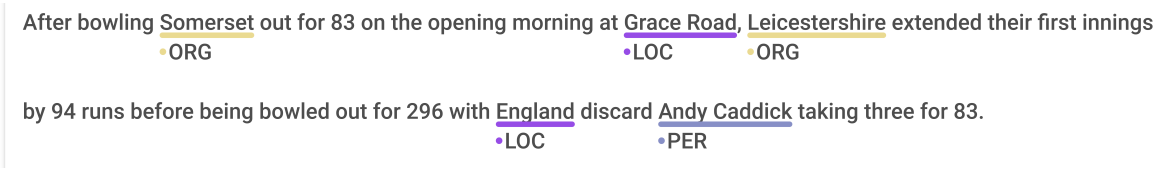


Figura 1. Ejemplo de reconocimiento de entidades nombradas con la herramienta de anotación *doccano*.[14]

tiempo, estas categorías han evolucionado a incluir conceptos más complejos en dominios especializados como la biomedicina.[12] Por ejemplo, en la Figura 1 se observa como ciertas palabras y frases han sido subrayadas y etiquetadas con abreviaciones para indicar la entidad que representan: Somerset y Leicestershire son organizaciones (etiqueta ORG), Grace Road y England son localizaciones (etiqueta LOC), y Andy Caddick es una persona (etiqueta PER). Puesto que a la hora de entrenar un modelo el texto debe llevar las etiquetas correspondientes, como se vio en el ejemplo anterior, existen herramientas de libre uso como *doccano*,[13] que facilitan la anotación manual de entidades, permitiendo la creación de conjuntos de datos.

NER tiene múltiples aplicaciones en varios dominios.[15] En tareas de recuperación de información, es crucial para identificar entidades relevantes tanto en las consultas de búsqueda como en los resultados, mejorando así la precisión. En la generación de resúmenes automáticos ayuda a resaltar la información más importante, mientras que en el monitoreo de redes sociales se emplea para detectar y clasificar menciones de marcas o temas de interés. En traducción automática incrementa la precisión

---

[12] Imed Keraghel, Stanislas Morbieu y Mohamed Nadif. *Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study*. 2024.

[13] Hiroki Nakayama et al. *doccano: Text Annotation Tool for Human*. Software available from https://github.com/doccano/doccano. 2018.

[15] Keraghel, Morbieu y Nadif, ver n. 12.

al preservar nombres y términos claves. En el ámbito de la salud, NER facilita la extracción de información sobre pacientes a partir de notas clínicas, literatura médica e historiales clínicos electrónicos, lo que mejora la gestión de la información médica. Finalmente, en la anonimización de documentos, la extracción de entidades o información sensible garantiza la privacidad de las personas y habilita el uso de datos delicados para la investigación sin comprometer la confidencialidad.

2.3. *TRANSFORMERS:* ARQUITECTURA ESTADO DEL ARTE EN NLP

Tradicionalmente, en modelos de NLP, como reconocimiento del habla o modelos de lenguaje grandes, se empleaban arquitecturas basadas en redes recurrentes, profundas o convolucionales.[16] Con el surgimiento de los *transformers*[17] se marcó un cambio fundamental, pues gracias a su mecanismo de *atención* logran capturar dependencias globales en los datos sin necesidad de recurrencia, lo que permite una mayor paralelización y un rendimiento significativamente superior. En la Figura 2 se ilustra la arquitectura de un *transformer*, la cual se basa en una estructura *encoder-decoder*: El *encoder* mapea una secuencia de representaciones de símbolos de entrada *inputs* (bien sean palabras o caracteres) a una secuencia de representaciones continuas llamadas *embeddings*. Como el modelo no tiene noción del orden de las palabras se le suma un *Positional Encoding* a cada embedding antes de pasar por el *encoder* y *decoder*. En el *decoder* se reciben dos entradas: la secuencia de salida generada hasta el momento desplazada un paso a la derecha y la salida del *encoder*. Como resultado, el *decoder* genera una secuencia de salida compuesta por las probabilidades de los símbolos, uno a la vez.

---

[16] Ahlawat, Aggarwal y Gupta, ver n. 6.

[17] Ashish Vaswani et al. “Attention is All you Need”. En: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

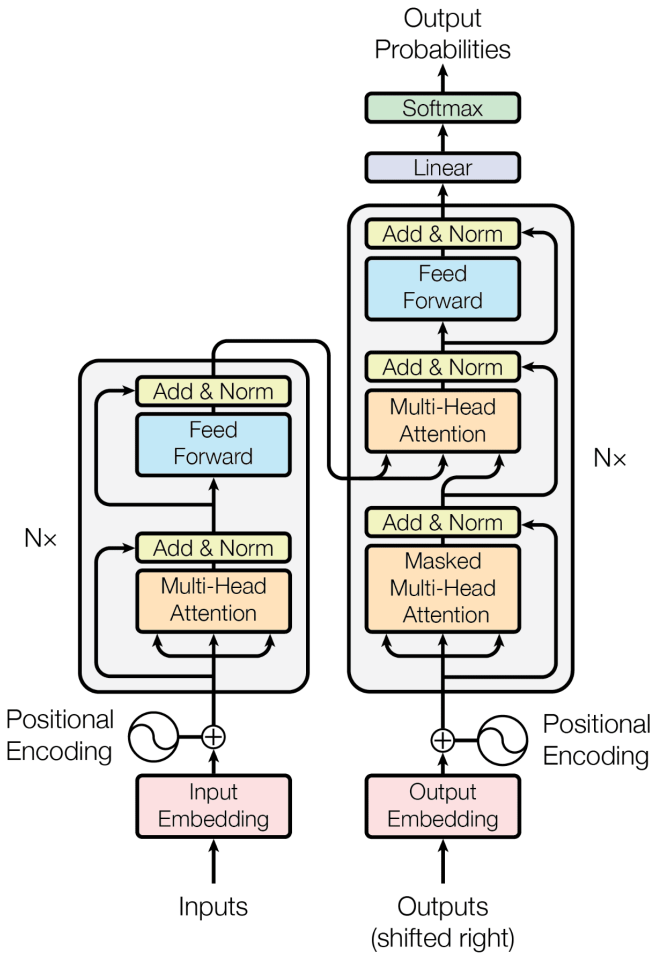


Figura 2. Arquitectura de un *transformer*.[18] La secuencia de entrada se procesa en el *encoder*, compuesto por capas de atención multi-cabeza y redes *feed-forward* que generan representaciones contextualizadas. El *decoder* recibe estas representaciones junto con la salida ya generada, aplicando atención tanto a la entrada como a la salida parcial para producir la siguiente palabra en la secuencia.

Como se mencionó previamente, y como se observa en la figura anterior, la novedad de los *transformers* se encuentra en su módulo de atención, que permite capturar dependencias a larga distancia dentro de los datos. Este modulo se puede describir

[18] Vaswani et al., ver n. 17

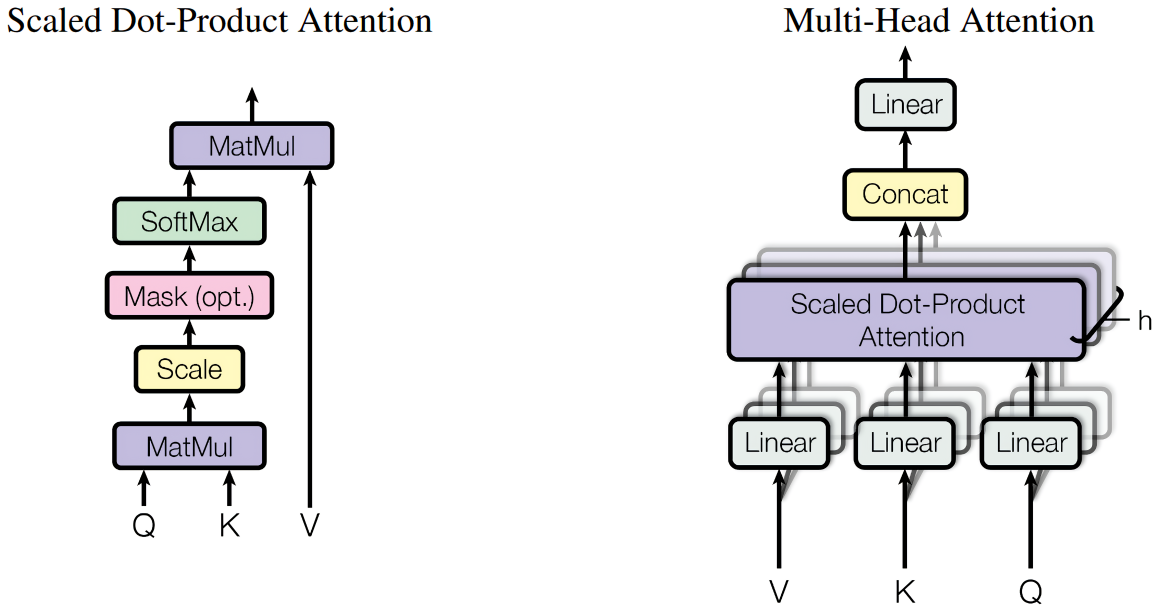


Figura 3. (Izquierda) *Scaled Dot-Product Attention*. (Derecha) *Multi-Head Attention*, consta de varias capas de atención que se ejecutan en paralelo.[19]

como una función que toma un vector consulta (*query*) y un conjunto de pares de vectores clave-valor (*key-value*), y produce una salida. La salida se calcula como una suma ponderada de los valores, donde el peso de cada uno se determina según su importancia para la consulta dada. En lugar de calcular la atención para una sola consulta, en la atención multi-cabeza (*multi-head attention*) se calculan múltiples atenciones de manera paralela, cada una utilizando su propia proyección de las matrices de consultas (Q), claves (K) y valores (V). Cada cabeza tiene su propio conjunto de proyecciones, lo que permite al modelo aprender diferentes representaciones de la entrada y enfocarse en múltiples aspectos de los datos de manera simultánea. En la Figura 3 se ilustran las operaciones de atención previamente explicadas.

---

[19] Vaswani et al., ver n. 17

2.4. RECONOCIMIENTO AUTOMÁTICO DEL HABLA (ASR)

El Reconocimiento Automático del Habla (ASR, por sus siglas en inglés *Automatic Speech Recognition*) es un subcampo interdisciplinario del NLP que permite el reconocimiento y la traducción del lenguaje hablado a texto.[20] Incorporar el lenguaje hablado en la interacción humano-máquina es esencial para la automatización de tareas; como el habla es la forma más natural y eficiente de comunicarnos, utilizarla como vía de entrada hace que la tecnología sea más accesible, fácil y cómoda de usar.[21]

Actualmente, los modelos de ASR se basan en una arquitectura de *Transformer* con *encoder-decoder* de extremo a extremo, la cual ha demostrado ser altamente efectiva y robusta para ASR, como se ve en Whisper,[22] ilustrado en la Figura 4. En este tipo de enfoques el proceso consiste en:

1. El modelo recibe una entrada de audio, esta se convierte en un *espectrograma Mel*, un tipo de espectrograma que utiliza una escala de frecuencias adaptada a la percepción humana del sonido, y se calcula tomando la magnitud logarítmica de las frecuencias a lo largo del tiempo.[24] Al convertir el audio en un espectrograma Mel se transforma en una representación visual más estructurada y

---

[20] Amarildo Rista y Arbana Kadriu. "Automatic Speech Recognition: A Comprehensive Survey". En: *SEEU Review* 15 (dic. de 2020), págs. 86-112. DOI: 10.2478/seeur-2020-0019.

[21] Mishaim Malik et al. "Automatic speech recognition: a survey". En: *Multimedia Tools and Applications* 80 (mar. de 2021), págs. 1-47. DOI: 10.1007/s11042-020-10073-7.

[22] Radford et al., ver n. 5.

[23] Tomada de OpenAI. *Introducing Whisper*. https://openai.com/index/whisper/. Último acceso: 16 de noviembre de 2025. 2022.

[24] Angel Mario Castro Martinez y Marc René Schädler. "Why do ASR Systems Despite Neural Nets Still Depend on Robust Features". En: *Interspeech 2016*. 2016, págs. 1883-1887. DOI: 10.21437/Interspeech.2016-1552.

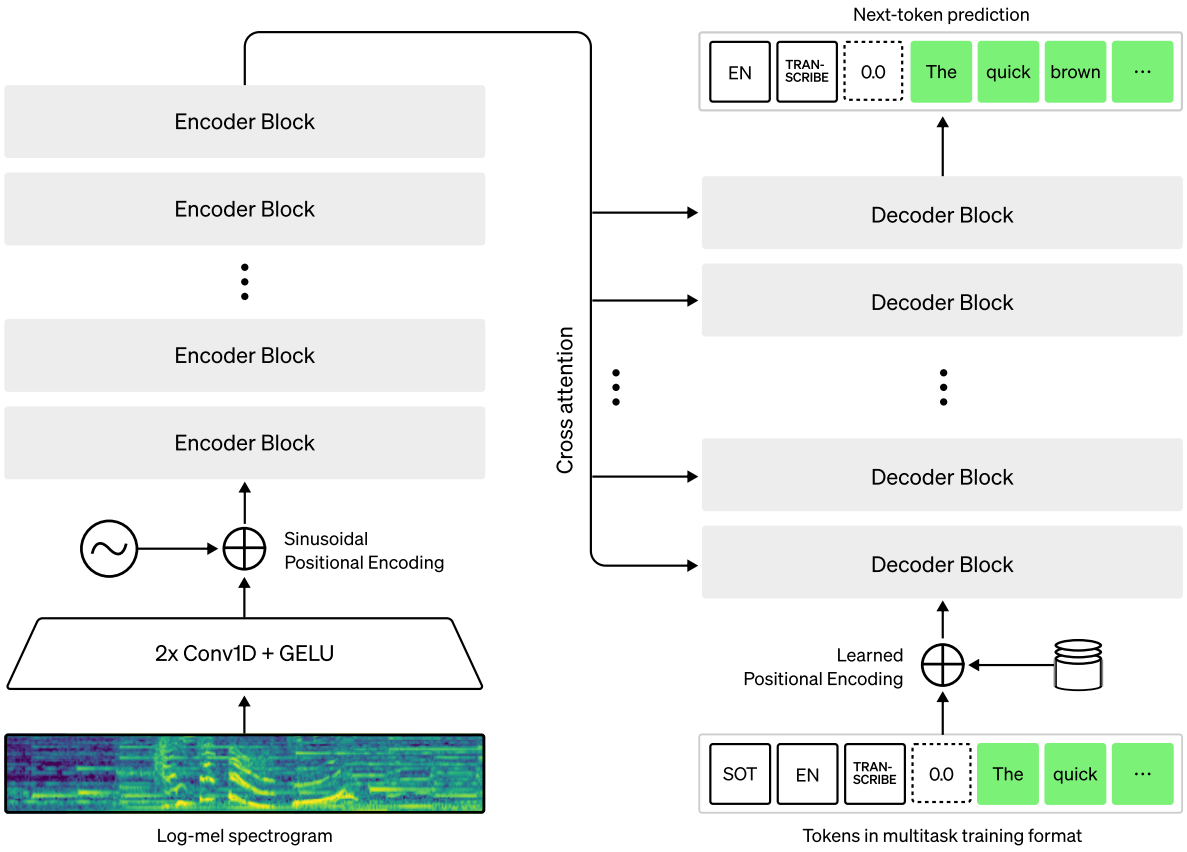


Figura 4. Descripción general de la arquitectura *Encoder-Decoder* de Whisper.[23] El audio de entrada se divide en fragmentos de 30 segundos y se convierte en un espectrograma log-Mel. Luego se pasa por el *encoder* y *decoder*. Este último se entrena para predecir la transcripción correspondiente, combinado con tokens especiales que dirigen al modelo para realizar tareas como la identificación del idioma, marcas de tiempo a nivel de frase, transcripción de voz multilingüe y la traducción de voz.

semejante a cómo la percibimos los humanos, permitiendo que el modelo ASR pueda interpretar patrones visuales y características más relevantes para el reconocimiento del habla.

2. Pre-procesamiento del audio: sus valores se normalizan globalmente para que tengan una media cercana a cero y se escalan al rango $[-1, 1]$.

3. El *encoder* toma esta representación del audio y la pasa por 2 capas convolucionales para extraer patrones locales en la señal. A la salida de la capa convolucional se le añaden *embeddings posicionales sinusoidales* para que el modelo tenga noción del orden temporal. A partir de ahí, la información pasa por varios bloques *encoder*, que combinan mecanismos de atención y capas *feed-forward* para encontrar relaciones complejas entre distintos segmentos de la señal. Cada bloque incluye conexiones llamados bloques residuales de preactivación (*pre-activation residual blocks*), que ayudan a que la información fluya mejor durante el entrenamiento, evitando que se degrade en redes profundas.

4. El *decoder* utiliza *embeddings posicionales aprendidos* durante el entrenamiento, las representaciones de tokens de entrada y salida compartidas, y el módulo de *atención* para generar el texto o las palabras que corresponden a ese audio. Además, puede condicionarse mediante *prompts*, bien sea palabras comunes o special tokens, que permiten orientar la decodificación hacia tareas específicas, como traducción u otras modalidades multitarea.

Aunque hoy en día existen varios modelos que han demostrado un rendimiento altamente competitivo en tareas de ASR, como Wav2Vec2, HuBERT, Whisper, y demás,[25] aún existe margen de mejora en diversos retos, desde la precisión de los sistemas hasta la inclusividad. Un desafío común en el reconocimiento de voz es la variabilidad de sílabas y fonemas de una misma palabra según la entonación y pronunciación del hablante, dificultando su identificación por parte de los modelos. Este problema se intensifica en contextos como la variación de acentos en diferentes lenguas o la diversidad acústica y lingüística típica del habla infantil. Así mismo, el *ruido* presente en las grabaciones, originado por el micrófono, el entorno, y otras variables, crea múltiples retos para su eliminación, convirtiéndose incluso en un tema

[25] Ahlawat, Aggarwal y Gupta, ver n. 6.

activo de investigación.[26]

Dentro del ámbito del ASR, también se incluyen tareas de post-procesamiento NLP para refinar las transcripciones resultantes y facilitar su uso en tareas posteriores. En este contexto, dos tareas fundamentales son: la anotación con marcas de tiempo y la diarización de hablantes. En la anotación con marcas de tiempo se busca asignar *timestamps* al inicio y final de palabras o frases, permitiendo aplicaciones prácticas como subtitulado automático temporizado, búsqueda de contenido del habla por palabras clave, o incluso facilitando otras tareas de habla como la conversación de voz a voz.[27] Por su parte, la existencia de las conversaciones entre múltiples interlocutores en podcasts, transmisiones, reuniones o videos resalta la necesidad de estimar automáticamente *¿quién habló cuándo?* a partir de una señal de audio grabada, tarea que realiza la diarización de hablantes.[28]

Junto con estas, otras tareas semánticas como el reconocimiento de entidades o el análisis de sentimientos suelen abordarse mediante arquitecturas modulares, en las que el ASR transcribe la voz a texto y su salida se procesa en etapas posteriores.[29] El problema de este enfoque es que presenta acumulación de errores: los errores de transcripción en la etapa de ASR se propagan a través de las distintas fases del flujo

---

[26] Sneha Basak et al. “Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems”. En: *Computer Modeling in Engineering Sciences* 135 (oct. de 2022), págs. 1-37. DOI: 10.32604/cmes.2022.021755.

[27] Ke Hu et al. “Word Level Timestamp Generation for Automatic Speech Recognition and Translation”. En: *Interspeech 2025*. 2025, págs. 2565-2569. DOI: 10.21437/Interspeech.2025-869.

[28] Douglas O’Shaughnessy. “Speaker Diarization: A Review of Objectives and Methods”. En: *Applied Sciences* 15.4 (2025). DOI: 10.3390/app15042002.

[29] Shashi Kumar et al. “TokenVerse: Towards Unifying Speech and NLP Tasks via Transducer-based ASR”. En: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, nov. de 2024, págs. 20988-20995. DOI: 10.18653/v1/2024.emnlp-main.1167.

de trabajo, lo que reduce el rendimiento de las tareas de NLP posteriores.[30] Para mitigar este efecto, un área de creciente interés es la integración de ASR con tareas de NLP en un mismo modelo. Ejemplos como WhisperNER permiten la transcripción conjunta del habla y el reconocimiento de entidades,[31] mientras que otros trabajos recientes exploran la fusión de ASR, diarización y asignación de marcas de tiempo en arquitecturas multimodales.[32,33,34] Estos enfoques muestran que el entrenamiento conjunto puede mejorar la coherencia en la asignación de hablantes y la anotación temporal; no obstante, suelen implicar *trade-offs* importantes, pues requieren un costo computacional elevado (Sortformer entrenado en 8 nodos con 8×Tesla V100, o DNCASR en una A100 de 80 GB), y en ocasiones sacrifican ligeramente la calidad del ASR a comparación de otros sistemas especializados.[35,36]

---

[30] Gil Ayache et al. "WhisperNER: Unified Open Named Entity and Speech Recognition". En: *arXiv preprint arXiv:2409.08107* (2024).

[31] Ayache et al., ver n. 30.

[32] Taejin Park et al. *Sortformer: A Novel Approach for Permutation-Resolved Speaker Supervision in Speech-to-Text Systems*. 2025. arXiv: 2409.06656 [eess.AS].

[33] Xianrui Zheng, Chao Zhang y Philip C. Woodland. *DNCASR: End-to-End Training for Speaker-Attributed ASR*. 2025. arXiv: 2506.01916 [eess.AS].

[34] Max Bain et al. "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio". En: *INTERSPEECH 2023* (2023).

[35] Laurent Shafey, Hagen Soltau e Izhak Shafran. "Joint Speech Recognition and Speaker Diarization via Sequence Transduction". En: sep. de 2019, págs. 396-400. DOI: 10.21437/Interspeech.2019-1943.

[36] Ke Hu et al. "Word Level Timestamp Generation for Automatic Speech Recognition and Translation". En: *Interspeech 2025*. 2025, págs. 2565-2569. DOI: 10.21437/Interspeech.2025-869.

2.5. WHISPERNER: RECONOCIMIENTO UNIFICADO Y ABIERTO DE ENTIDADES NOMBRADAS Y DEL HABLA

WhisperNER es un modelo unificado que realiza simultáneamente la transcripción del habla y el reconocimiento de entidades, basado en la arquitectura del modelo Whisper.[37] En su enfoque condicionan el proceso de decodificación de Whisper a un conjunto de etiquetas de entidad $t = [t_1, t_2, \ldots, t_k]$ para el que cada t_i representa un tipo de entidad específica, como persona, localización, etc. El flujo del proceso sigue siendo el mismo que la arquitectura nativa de Whisper: el *decoder* genera cada token y_t basándose en los tokens anteriores y los estados ocultos del *encoder* $t = \text{Decoder}(y_{1:t-1}, h, t)$. Su salida es una secuencia de tokens $y = [y_1, y_2, ..., y_n]$, que comprende tanto el texto transcrito como las etiquetas de entidad correspondientes, como se ilustra en la Figura 5. El modelo está entrenado para minimizar la *standard cross-entropy loss* entre la secuencia de salida predicha y y la secuencia *ground truth* y^*, que incluye tanto la transcripción correcta como las etiquetas de entidad correctas, véase la Ecuación (1)

$$\mathcal{L}(y, y^*) = -\sum_{t=1}^{n} \log P(y_t = y_t^* \mid y_{1:t-1}, h, t)\,. \tag{1}$$

2.6. WHISPERX: TRANSCRIPCIÓN TEMPORALMENTE PRECISA DE AUDIOS DE LARGA DURACIÓN

A pesar de que Whisper ha demostrado resultados impresionantes de ASR en diferentes dominios e idiomas, las marcas de tiempo predichas a nivel de enunciado

---

[37] Ayache et al., ver n. 30.

[38] Ayache et al., ver n. 30

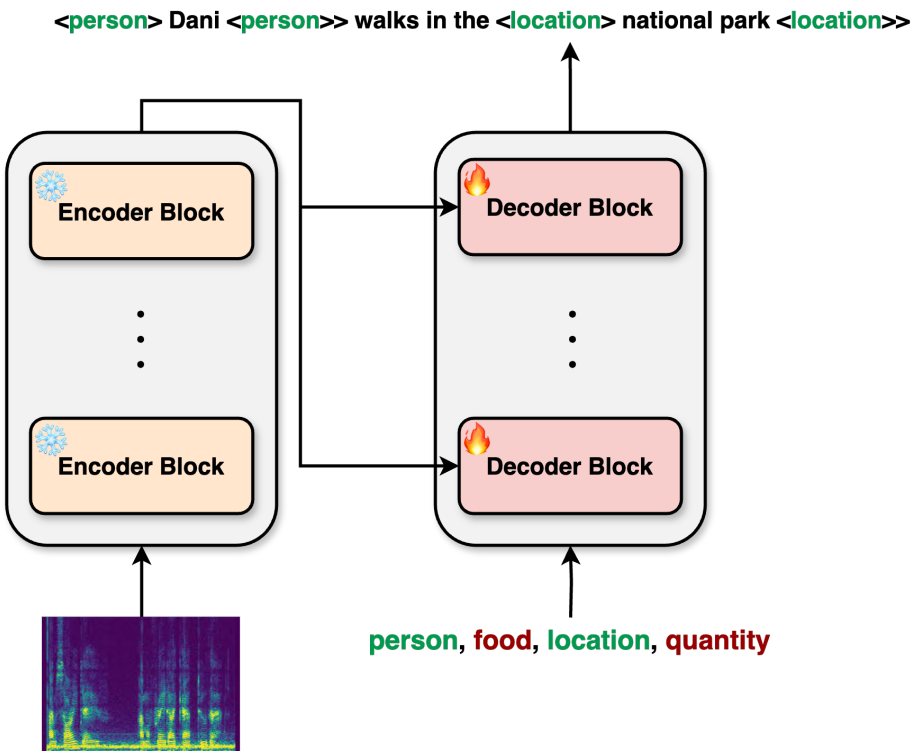


Figura 5. Arquitectura de WhisperNER. Se proporciona un conjunto de tipos de entidad como *prompt* para el *decoder*. Durante el entrenamiento se proporcionan entidades positivas (en color verde) y negativas (en color rojo). Al momento de la inferencia, el modelo puede generalizar a nuevos tipos de entidad no observados durante el entrenamiento.[38]

suelen presentar imprecisiones, y no se ofrecen de manera nativa a nivel de palabra *out-of-the-box*. Debido a este problema surge WhisperX, un sistema ASR con precisión temporal y marcas de tiempo a nivel de palabra que utiliza la detección de la actividad vocal (VAD, por sus siglas en inglés *Voice Activity Detection*) y la alineación forzada de fonemas, demostrando un rendimiento de vanguardia en la transcripción y anotación temporal de audios.[39] Tiene 4 componentes, como se muestran en la Figura 6, y pueden resumirse como sigue:

[39] Bain et al., ver n. 34.

- **Voice Activity Detection (VAD):** VAD se refiere al proceso de identificar regiones dentro de un flujo de audio que contiene voz. Para WhisperX se presegmenta el audio de entrada con VAD para (1) evitar *forward-passes* innecesarios de ASR en largas regiones de audio sin voz; (2) dividir el audio en fragmentos cuyos cortes no caen en regiones con voz, y así reducir errores en los bordes y facilitar la transcripción en paralelo; (3) acotar la alineación a nivel de palabra a segmentos más pequeños y confiables, evitando depender de las marcas de tiempo de Whisper.

- **VAD Cut & Merge:** Como los segmentos de voz obtenidos en el paso anterior pueden tener una longitud arbitraria mucho mayor o menor que la duración máxima de entrada del modelo ASR (en este caso, de Whisper), los segmentos más largos no pueden transcribirse con un solo *forward-pass* y los más cortos pierden contexto y aumentan el tiempo de inferencia al necesitar varios *forward-passes*. Para solucionarlo, cortan segmentos demasiado largos en partes manejables, eligiendo los cortes en zonas con baja VAD (Min-cut) y unen segmentos demasiado cortos para recuperar contexto y reducir cómputo (Merge).

- **Whisper Transcription:** Los segmentos resultantes son los que serán transcritos eficientemente en paralelo con Whisper sin *causal conditioning*, es decir, sin usar el texto previo para predecir el siguiente token.

- **Forced Phoneme Alignment:** Para cada segmento de audio s_i con su transcripción T_i, compuesta por una secuencia de palabras $T_i = [w_0, w_1, ..., w_m]$, se busca estimar el tiempo de inicio y fin de cada palabra. Ello se hace con un modelo de reconocimiento de fonemas que toma como entrada un segmento de audio s_i y genera una matriz de logits $L \in \mathbb{R}^{K \times T}$, siendo K el número de phonemas posibles que el modelo puede predecir y T el número de pasos de

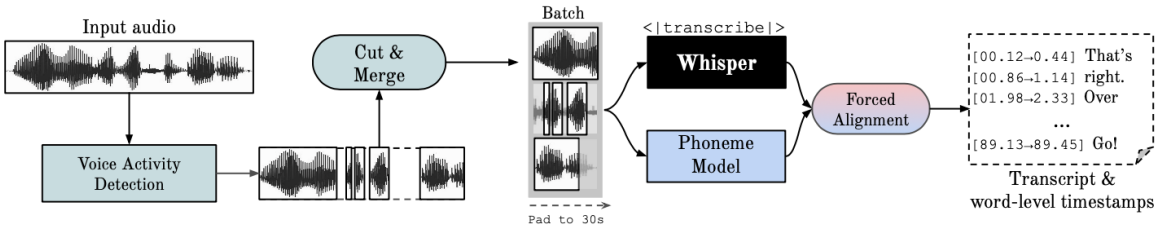


Figura 6. Arquitectura de WhisperX. El audio de entrada se segmenta primero con VAD y luego se corta y fusiona en fragmentos de entrada de aproximadamente 30 segundos. Los fragmentos resultantes se transcriben en paralelo con Whisper y se alinean forzosamente con un modelo de reconocimiento de fonemas para producir marcas de tiempo precisas a nivel de palabra.[40]

tiempo (frames) en que se divide el audio. El valor $L_{k,t}$ indica la probabilidad de que el phonema k ocurra en el tiempo t.

2.7. LORA: ADAPTACIÓN DE BAJO RANGO DE MODELOS DE LENGUAJE GRANDES

A medida que se pre-entrenan modelos más grandes, un *fine-tuning* completo, que reentrena todos los parámetros del modelo, se vuelve menos factible. Tomando como ejemplo GPT-3 de 175 billones de parámetros, desplegar instancias independientes de modelos afinados, cada uno con 175B de parámetros, resulta extremadamente costoso. Como solución a estas limitaciones surge *LoRA* (Low-Rank Adaptation of Large Language Models)[41]; esta técnica propone una alternativa al ajuste completo de parámetros a la hora de entrenar modelos. Para cada capa i se congela la matriz de pesos $W_i \in \mathbb{R}^{d\times k}$ y se inyectan dos matrices entrenables $B \in \mathbb{R}^{d\times r}$ y $A \in \mathbb{R}^{r\times k}$, con $r \ll \min(d, k)$, representando la matriz de pesos mediante una descomposición

[40] Ayache et al., ver n. 30

[41] Edward J Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". En: *International Conference on Learning Representations*. 2022.

de bajo rango en la Ecuación (2):

$$W_i + \Delta W = W_i + BA. \tag{2}$$

Siendo $h = W_0 x$ la salida de cada capa (*hidden state*) original, la nueva salida se calcula como la Ecuación (3):

$$h = W_0 x + \Delta W x = W_0 x + BAx. \tag{3}$$

Hu et al.[42] reportan en sus resultados que LoRA no solo se mantiene competitivo, sino que puede superar al *fine-tuning* completo incluso en modelos de escala extrema como GPT-3. Esto confirma que su diseño permite alcanzar altos niveles de desempeño entrenando una fracción mínima de los parámetros totales, logrando una relación mucho más eficiente entre costo y rendimiento.

[42] Hu et al., ver n. 41.

3. MÉTODO PROPUESTO

En este trabajo se desarrolló un modelo de lenguaje natural para la transcripción anonimizada de grabaciones de audio del español de Colombia, incorporando tareas *downstream* pertinentes a NLP como diarización de hablantes y anotación con marcas de tiempo. El enfoque propuesto se basa en la arquitectura tipo *transformer encoder-decoder* usada por WhisperNER[43] y el *pipeline* propuesto por WhisperX,[44] véase la Figura 7.

3.1. FLUJO DE TRABAJO (*PIPELINE*) PARA LA TRANSCRIPCIÓN ANONIMIZADA

Para nuestro método, las entradas son un archivo de audio y una lista de *prompts* textuales $p = \{p_1, p_2, ..., p_k\}$, para el cual cada p_i representa un tipo de entidad en específico. El modelo WhisperNER se adapta a nuevas entidades pero para este proyecto por defecto se trabaja con 4 tipos de entidades: **"person"**, **"location"**, **"organization"** y **"miscellaneous"**. Antes de procesar el audio con el modelo de ASR se usa un modelo VAD -bien sea *Silero*,[45] un modelo ligero en PyTorch optimizado para detección rápida, o *pyannote*,[46] modelo más complejo con mayor robustez en escenarios de diarización- para pre-segmentarlo en regiones que contienen

---

[43] Ayache et al., ver n. 30.

[44] Bain et al., ver n. 34.

[45] Silero Team. *Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. `https://github.com/snakers4/silero-vad`. 2024.

[46] Hervé Bredin et al. "pyannote.audio: neural building blocks for speaker diarization". En: *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*. Barcelona, Spain, mayo de 2020.

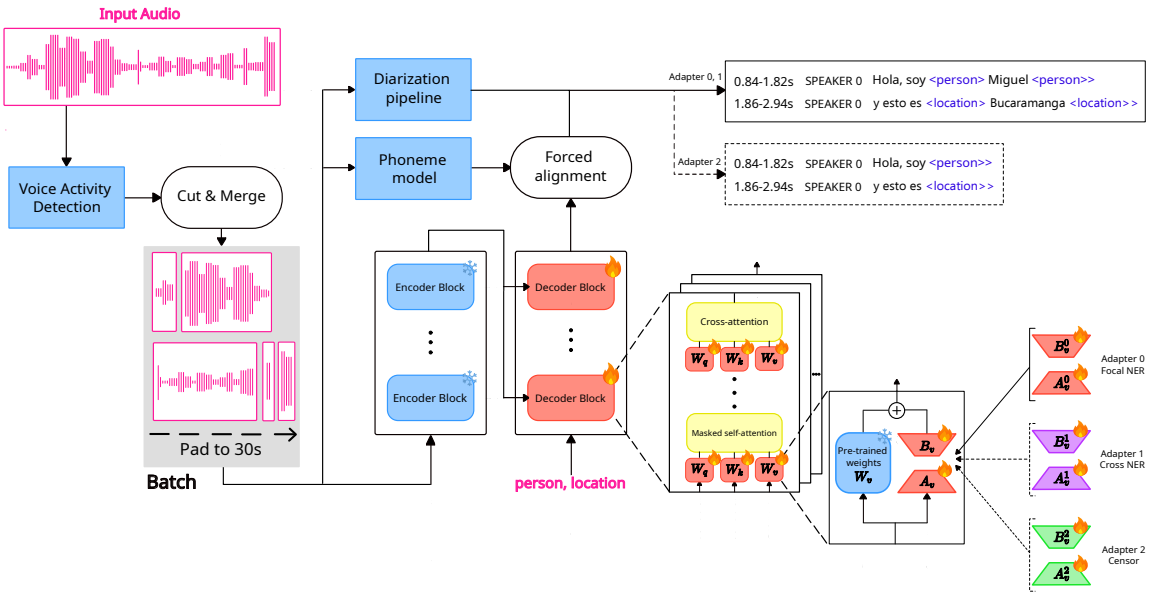


Figura 7. Flujo de trabajo del método propuesto para la transcripción anonimizada de grabaciones de audio del español de Colombia, con anotación temporal y diarización de hablantes. El modelo integra el sistema propuesto por WhisperX para obtener marcas temporales precisas, complementado con la *pipeline* de diarización de pyannote. Para la transcripción, se emplea un *fine-tuning* de WhisperNER utilizando LoRA sobre los pesos de atención en el *decoder*. Durante la inferencia, es posible seleccionar distintos *adapters* LoRA en función del modelo deseado (Focal NER, Cross NER o censura).

habla activa. Para esta tarea, la forma de onda del audio de entrada se representa como una secuencia de vectores de características acústicas extraídos por paso de tiempo $A = \{a_1, a_2, ..., a_T\}$ y la salida es una secuencia de etiquetas binarias $y = \{y_1, y_2, ..., y_T\}$, para la cual $y_t = 1$ significa que hay habla en el paso de tiempo t, y $y_t = 0$ viceversa. Estas predicciones son luego representadas como una secuencia de segmentos de habla activos $s = \{s_1, s_2, ..., s_N\}$, con índices de inicio y fin $s_i = (t_0^i, t_1^i)$. Luego, se hace una operación *Cut & Merge* (cortar y fusionar en español): cada segmento s de duración $T > 0$ se corta en el punto con puntuación mínima de activación de voz; este corte se restringe entre $\frac{1}{2}|A_{train}|$ y $|A_{train}|$, siendo $|A_{train}|$ la duración máxima que acepta el modelo ASR durante entrenamiento, en nuestro caso

30 segundos. En caso que hayan fragmentos s muy cortos, se hace la operación inversa; teniendo 2 segmentos adyacentes s_i y s_{i+1}, si la duración total del intervalo combinado $d_{i,i+1} = t_1^{i+1} - t_0^i$ es menor a un umbral de duración máxima τ, para el que $\tau \leq |A_{train}|$, entonces se fusionan.

Los segmentos de voz resultantes, con una duración temporal (en segundos) aproximadamente igual a la longitud promedio de los ejemplos de entrenamiento del modelo ASR, $|s_i| \approx |A_{\text{train}}|$, son luego preprocesados para convertirlos en una representación adecuada para el modelo. Este preprocesamiento sigue el pipeline estándar de Whisper, transformando la señal en un espectrograma Log-Mel $X \in \mathbb{R}^{T \times 80}$, calculado con ventanas de 25 ms y un *stride* de 10 ms, siendo T el número de frames de tiempo generados a partir de la duración del audio y estos parámetros. Nuestro X resultante representa las características de cada audio de entrada que entran al *encoder*, el cual produce una secuencia de *hidden states*, $h = Encoder(X)$, que se usan para condicionar el *decoder* junto a nuestro set de etiquetas de entidad p. El decoder genera una secuencia de tokens $y = [y_1, y_2, ..., y_n]$, que comprende tanto el texto transcrito como las etiquetas de entidad correspondientes, por ejemplo: *Hola, soy <person> Miguel <person>>*. Para ASR se tienen tres *adapters* LoRA distintos, cada uno más competente en una tarea específica: Focal NER ofrece mejor rendimiento en transcripción (ASR), Cross NER en reconocimiento de entidades (NER), y censura, aunque no supera a los anteriores en esas tareas, incorpora censura directamente en el proceso de inferencia, eliminando la necesidad de un módulo adicional. Esto permite al usuario seleccionar el *adapter* más adecuado según sus necesidades.

Para cada segmento de audio s_i usamos un modelo de reconocimiento de fonemas (*Phoneme model* en la Figura 7), definiendo fonemas como las unidades mínimas del habla (ejemplo, el sonido /k/ en *casa*), para obtener una matriz de *logits* $L \in \mathbb{R}^{K \times T}$, siendo K el número de clases de fonemas que el modelo sabe reconocer y T el número de frames de audio. Con la transcripción asociada y_i, se obtiene su

representación en fonemas mediante el diccionario del modelo. Luego, estos fonemas de referencia se alinean con las probabilidades extraídas del audio utilizando *Dynamic Time Warping (DTW)* (bloque *Forced alignment* en la Figura 7), lo que permite encontrar la trayectoria temporal más probable de cada fonema. Finalmente, las marcas de inicio y fin de cada palabra se determinan a partir de los límites de sus fonemas correspondientes. Para el caso del español se usa un modelo wav2vec 2.0,[47] arquitectura auto-supervisada para el aprendizaje de representaciones del habla a partir de audio crudo, pre-entrenado con 10.000 horas de audio sin etiquetar del conjunto de datos *VoxPopuli* y ajustado para ASR en 166 horas de audio transcrito del subconjunto en español.[48]

Para la diarización de hablantes se usa el *diarization pipeline* de *pyannote.audio*,[49] por su desempeño competitivo frente al estado del arte y su facilidad de uso *plug-and-play*. El *pipeline* divide el audio en segmentos con VAD, extrae *embeddings* de cada segmento de audio y los agrupa para decidir qué segmentos pertenecen al mismo hablante. Luego, para cada segmento o palabra de la transcripción se calcula cuánto se solapa en el tiempo con cada segmento de diarización, y si hay solapamiento se asigna como hablante el que más tiempo comparta.

---

47 Alexei Baevski et al. “wav2vec 2.0: a framework for self-supervised learning of speech representations”. En: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Vancouver, BC, Canada: Curran Associates Inc., 2020.

48 Changhan Wang et al. “VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation”. En: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, ago. de 2021, págs. 993-1003.

49 Bredin et al., ver n. 46.

3.2. CREACIÓN DE BASE DE DATOS

Frente a la ausencia de conjuntos de datos públicos anotados para Colombia, se procedió a crear un dataset por medio de un *pipeline* automatizado que combina *web scraping* y descarga de audios. En primer lugar, se define una función para obtener el audio de un video a partir de su URL con *yt-dlp*,[50] una herramienta para descargar videos y audios de múltiples sitios web de streaming, utilizando un postprocesador con *FFmpeg*,[51] una biblioteca para el procesamiento de audio y video, para extraer el audio y convertirlo a formato `wav` con calidad de 192 kbps. Para la recolección de los enlaces de los videos se usa *Selenium*[52] como driver de navegador, accediendo a la página de TikTok mediante una palabra clave o etiqueta específica (por ejemplo, #Bucaramanga). Una vez cargado el contenido principal, se realiza un scroll iterativo hasta que no se detecta contenido nuevo y se extraen todos los enlaces que contienen videos, filtrando aquellos que no corresponden a material audiovisual. Finalmente, cada enlace de video se procesa con la función de descarga de audio, registrando el progreso y los posibles errores que se presenten durante el proceso. Este flujo permite automatizar la creación de un dataset consistente y de alta calidad, partiendo de búsquedas dinámicas en TikTok y obteniendo archivos de audio listos para su uso.

Para asegurar diversidad de acentos y expresiones en el dataset, se escogieron 13 regiones colombianas que representan distintos dialectos y acentos descritos en el Atlas Lingüístico-Etnográfico de Colombia (ALEC) y otros estudios.[53] Aunque

---

[50] yt-dlp Project. *yt-dlp: A feature-rich command-line audio/video downloader*. `https://github.com/yt-dlp/yt-dlp`. 2025.

[51] FFmpeg Developers. *FFmpeg: multimedia framework*. `https://ffmpeg.org/`. 2025.

[52] SeleniumHQ. *Selenium WebDriver*. `https://www.selenium.dev/`. 2025.

[53] **ALEC**.

Ciudad	Región	Dialecto	# Audios	Duración (min)
Arauca	Arauca	Llanero	184	62.93
Barranquilla	Atlántico	Costeño – Atlántico	521	186.29
Bogotá	Cundinamarca	Andino – Oriental	415	149.36
Bucaramanga	Santander	Andino – Oriental	466	163.10
Cali	Valle del Cauca	Andino – Occidental	379	133.15
Cúcuta	Norte de Santander	Andino – Oriental	422	150.64
Medellín	Antioquia	Andino – Occidental	467	152.51
Neiva	Huila	Andino – Oriental	279	96.58
Pasto	Nariño	Andino – Occidental	330	114.09
Quibdó	Chocó	Costeño – Pacífico	219	73.96
San Andrés	-	-	219	78.98
Tunja	Boyacá	Andino – Oriental	342	118.05
Yopal	Casanare	Llanero	279	95.93
Total	-	-	**4522**	**1575.57**

Cuadro 1. Total de audios recopilados, comparados por ciudad, región, dialecto, cantidad y duración.

no cubren toda la riqueza dialectal del país, incluyen subdialectos principales como el acento *paisa*, y garantizan una muestra diversa de los 5 dialéctos reconocidos: costeño atlántico, costeño pacífico, andino occidental, andino oriental y llanero, además del español isleño hablado en San Andrés, el cual no está dentro de la clasificación. La recolección de datos se realizó en *TikTok*, aprovechando su sistema de *hashtags*; esta elección se fundamenta en que los usuarios tienden a etiquetar la ciudad en sus publicaciones, lo que facilita la identificación de videos asociado a cada región y favorece la representatividad del corpus. En el Cuadro 1 se muestra la ciudad, región, dialecto, número de audios recopilados y duración total del conjunto de datos.

Tras recopilar los audios de cada ciudad, se generaron transcripciones automáticas iniciales utilizando **Whisper large-v2**.[54] Posteriormente, se verificó cada audio y su transcripción manualmente, descartando aquellos inutilizables, por ejemplo en otros

[54] Radford et al., ver n. 5.

idiomas, con acentos de regiones distintas a la esperada o con sólo música. Debido a este descarte se presenta un desbalance de clases en la base de datos final. Adicionalmente, la propia plataforma de *TikTok* limitó el número de videos accesibles por cada etiqueta. En algunos casos el sistema permitía descargar alrededor de 400 videos, en otros hasta 700, pero siempre existía un punto en el que dejaba de mostrar nuevos resultados. Esto generó diferencias notorias entre regiones: ciudades con mayor volumen de publicaciones, como Bogotá, aportaron muchos más audios que regiones con menor actividad o visibilidad, como Pasto.

Finalmente, para obtener las etiquetas de entidades de cada transcripción se utilizó el modelo **NER-spanish-large** de la librería Flair,[55] basado en la arquitectura FLERT para reconocimiento de entidades nombradas, por sus resultados competitivos en el dataset CoNLL-03 (Spanish) con un F1-score del 90,54 %. Este modelo sólo etiqueta 4 tipos de entidades: **"PER"** (nombre de una persona), **"LOC"** (nombre de una localización), **"ORG"** (nombre de una organización), **"MISC"** (otro nombre, entidades misceláneas). En el caso de las etiquetas NER, no fue posible realizar una verificación exhaustiva debido a la complejidad y tiempo que requería esta tarea. El total de entidades recopiladas se ve en el Cuadro 2 junto a su split de entrenamiento, validación y test.

Etiqueta de entidad	Total	Train	Validation	Test
PER	1447	903	217	327
ORG	830	481	123	226
LOC	4610	2693	932	985
MISC	2126	1276	425	422
Total	**9013**	**5353**	**1697**	**1961**

Cuadro 2. Total de entidades en la base de datos.

Para manejar el desbalance de clases entre ciudades, se optó por un *split* de

[55] Stefan Schweter y Alan Akbik. *FLERT: Document-Level Features for Named Entity Recognition.* 2020. arXiv: `2011.06993 [cs.CL]`.

entrenamiento, validación y test que (1) distribuyera de la manera más equilibrada posible las ciudades en cada subconjunto y (2) garantizara que, en los casos donde un audio se dividía en varios segmentos según su duración, todos los segmentos pertenecieran a un mismo subconjunto, evitando así fugas de información (*data leakage*). La Figura 8 muestra un gráfico de barras que representa la distribución de los registros del dataset según la ciudad y partición de los datos (entrenamiento, validación y test), donde puede observarse el claro desbalance de clases entre ciudades.

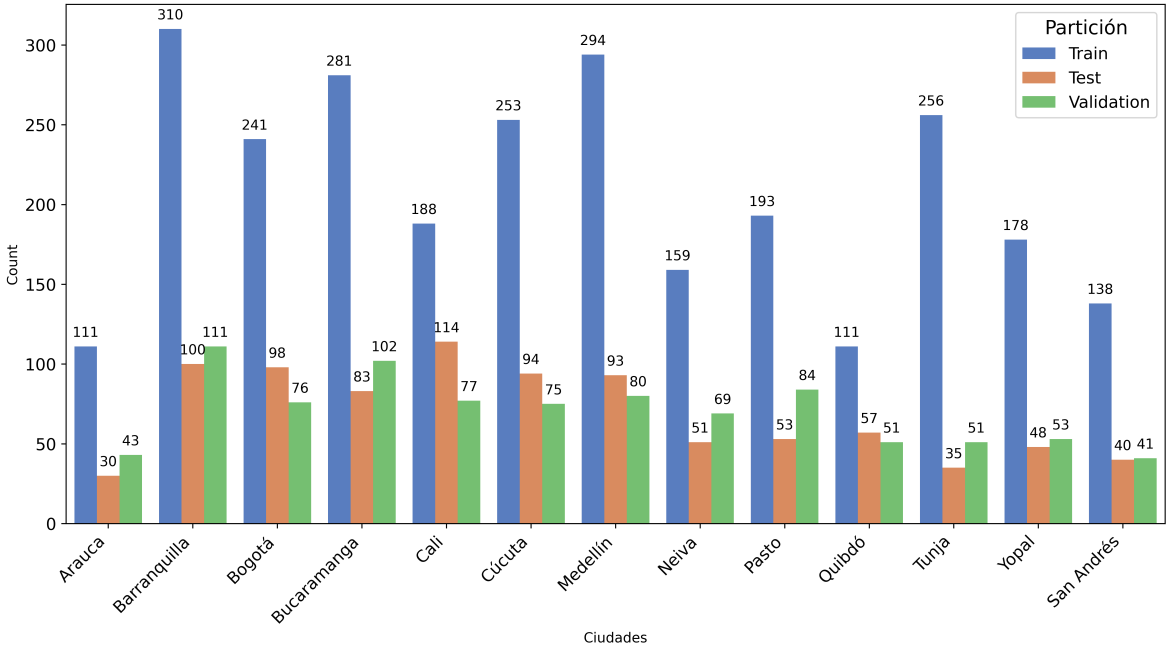


Figura 8. Distribución de las ciudades en los splits de train, validation y test.

---

55 Ayache et al., ver n. 30

3.3. ESTRATEGIA DE ENTRENAMIENTO

El entrenamiento del modelo ASR consistió en un *fine-tuning* sobre el *decoder* de WhisperNER con *LoRA*[56] durante 1000 pasos, entiéndase *paso* como una actualización de parámetros del modelo, con evaluación en el conjunto de validación cada 40 pasos, empleando la función de pérdida como métrica principal de selección de *checkpoints*. Los *adapters* LoRA se entrenaron congelando los parámetros del modelo base y actualizando únicamente las proyecciones de *query*, *key* y *value* de atención (*cross-attention* y *self-attention*) del *decoder*, y su configuración consistió en un rango de 16, un factor de escalado *alpha* de 16 y un dropout del 10 %; véase en la Figura 7 los *adapters* entrenados. Todos los experimentos se llevaron a cabo en una GPU Tesla P100, utilizando un tamaño de lote *(batch size)* de 4 con *gradient accumulation* con 8 pasos para simular un tamaño de lote mayor sin superar la memoria GPU disponible; esta elección se debe a que en modelos grandes como Whisper, el uso de lotes más grandes reduce la varianza del gradiente y tiende a estabilizar y mejorar el entrenamiento. Dado al límite de cómputo que impone la GPU, la estrategia adoptada fue acumular gradientes durante varios mini-batches antes de realizar la actualización de parámetros. El entrenamiento se realizó utilizando el optimizador *prodigy*,[57] el cual ajusta dinámicamente el tamaño de paso sin necesidad de un tuning manual extenso de sus hyperparámetros, como sería el caso de AdamW, convergiendo a un mínimo mucho más rápido. Así mismo, se usó un programador de tasa de aprendizaje de tipo coseno *(cosine scheduler)* para un ajuste dinámico de la tasa de aprendizaje (LR, por sus siglas en inglés *Learning Rate*) durante el

---

[56] Hu et al., ver n. 41.

[57] Konstantin Mishchenko y Aaron Defazio. “Prodigy: An Expeditiously Adaptive Parameter-Free Learner”. En: *Proceedings of the 41st International Conference on Machine Learning*. Ed. por Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, 21–27 Jul de 2024, págs. 35779-35804.

entrenamiento.

Se hicieron pruebas usando *cross-entropy loss*, considerada como la función de pérdida estándar en ASR, y *focal loss*, propuesta como una alternativa para NER que produce una ligera mejora en el rendimiento del modelo para casos donde hay desbalance de clases.[58] El término $\log P(y_t = y_t^* \mid y_{1:t-1}, h, t)$ en la Ecuación (1) se encarga de maximizar la probabilidad de la clase correcta en cada posición de la secuencia y penalizar las predicciones erróneas; sin embargo, como todas las muestras mal clasificadas le pesan lo mismo, no distingue entre ejemplos muy difíciles o casi correctos, así que en casos de desbalance de clases los ejemplos fáciles y frecuentes terminan dominando. Aquí es cuando entra la *focal loss* en la Ecuación (4), que usa 2 términos para abordar este problema: (1) un factor modulador $(1 - p_t)^\gamma$ que reduce la contribución de los ejemplos fáciles y concentra la optimización en los difíciles y (2) un factor de ponderación α_t que compensa este desbalance.

$$\mathcal{L}_{\mathsf{FL}}(p_t) = -\alpha_t \, (1 - p_t)^\gamma \, \log(p_t). \tag{4}$$

Es importante señalar que la anonimización de entidades puede abordarse de dos maneras. La primera consiste en emplear uno de los *adapters* entrenados para NER, cuya salida incluye etiquetas como *<person> Miguel <person>>*. Posteriormente, un módulo de post-procesamiento elimina la información sensible y conserva únicamente la etiqueta genérica, en este caso *<person>>*. La segunda alternativa es utilizar un *adapter* que aplica la censura directamente durante el *forward-pass*, de modo que la salida del modelo ya contiene únicamente la etiqueta, por ejemplo: *Hola, soy <person>>*. Ambas soluciones están disponibles en el método propuesto y su elección

---

[58] Zhiqiang Huang, Liang He, Yu Yang et al. “Application of machine reading comprehension techniques for named entity recognition in materials science”. En: *Journal of Cheminformatics* 16.76 (2024). DOI: 10.1186/s13321-024-00874-5.

depende del nivel de precisión requerido; en la sección de resultados se presentan las métricas correspondientes a cada enfoque.

4. RESULTADOS

4.1. CARACTERÍSTICAS DE LA BASE DE DATOS PROPUESTA

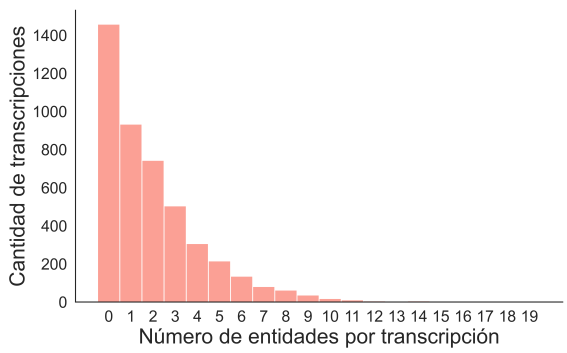
Frente a la escasez de conjuntos de datos que proporcionen audios colombianos junto a su respectiva transcripción se usó un único conjunto de datos recopilado por *web scrapping*, como se detalló en la Sección 4, como entrenamiento y *benchmark* para todas las pruebas y evaluaciones. Cada partición es representada en un archivo en formato JSON, todos los audios se dividieron en segmentos de 30 segundos o menos según VAD, y están almacenados por carpetas según su ciudad. Cada registro del conjunto de datos tiene 6 campos:

- **segment_id:** Identificador único del audio, con formato `tiktok_<ciudad>_<id_audio>_<id_segmento>`.
- **audio_path:** Ruta relativa hacia cada archivo de audio.
- **transcription:** Transcripción correspondiente a cada segmento.
- **quality:** Nivel de calidad del audio (se determinó cualitativamente)
  - `0`: Audio muy bueno o sin problemas.
  - `1`: Calidad regular; algunas expresiones no se entienden.
  - `2`: Muy mala calidad, casi inaudible o incomprensible.
- **region:** Ciudad correspondiente a cada audio.
- **entities:** Lista con todas las entidades anotadas en la transcripción, en formato *span*: `[carácter inicial, carácter final, entidad, palabra, descripción de la palabra]`. Este último campo es opcional.

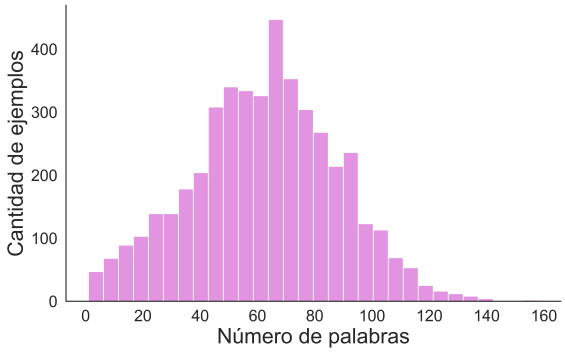
Un ejemplo de cómo se ven estos JSON es:


```
[
{
  "segment_id": "tiktok_Bucaramanga_103_5",
  "audio_path": "bucaramanga/tiktok_Bucaramanga_103_5.wav",
  "transcription": "y a Andrés Ferney Contreras Torres, aguante la banda de
  ↪  Wilson Díaz Orduz. Oiga, por ahí me dijeron que usted se va a bailar
  ↪  cumbia, ¿de verdad? Sí, yo las sé bailar poporro. ¿Vamos a bailar o
  ↪  qué? Vamos a bailarlas. Hágale.",
  "quality": 0,
  "region": "Bucaramanga",
  "entities": [
    [5, 35, "PERSON", "Andrés Ferney Contreras Torres", "Person; a named
    ↪  individual human being"],
    [57, 74, "PERSON", "Wilson Díaz Orduz", "Person; a named individual human
    ↪  being"]
  ...
},
...
]
```

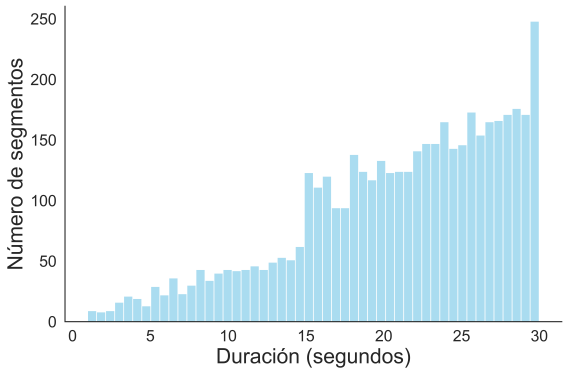

En la Figura 9 se presenta un análisis exploratorio del conjunto de datos utilizado, la cual muestra tres aspectos clave del corpus. En primer lugar, se observa que la mayoría de las transcripciones contienen pocas o ninguna entidad reconocida, lo que evidencia una distribución altamente desbalanceada para tareas NER. En segundo lugar, la longitud de las transcripciones (medida en número de palabras) presenta una distribución similar a la normal, con una media en torno a las 60-80 palabras. Finalmente, se analiza la duración de los segmentos de audio, donde se aprecia una



(a) Número de entidades por transcripción



(b) Longitudes de transcripciones



(c) Duraciones de segmentos

Figura 9. Distribuciones del conjunto de datos: entidades reconocidas, longitudes de transcripción y duración de audio por registro.

concentración creciente hacia el límite superior de 30 segundos, lo que indica que la mayoría de tiktoks superaban la duración máxima.

4.2. MÉTRICAS DE EVALUACIÓN

Para evaluar el rendimiento del método propuesto se adoptaron las métricas estándar utilizadas en el estado del arte para las tareas de ASR y NER.

∗ **Word Error Rate (WER) y Character Error Rate (CER):** El rendimiento de un ASR es difícil de calcular puesto a que su salida puede no tener la misma longitud que el

ground truth. El WER es una métrica comúnmente utilizada ya que calcula el error a nivel de palabra en lugar de a nivel de fonema.[59] El CER da un análisis mucho más fino que el WER y se formula de la misma manera pero a nivel de carácter. Su ecuación es:

$$\text{WER}, \text{CER} = \frac{S + D + I}{N}, \tag{5}$$

donde S es el número de sustituciones realizadas en el texto de salida comparado con el *ground truth*, D es el número de eliminaciones realizadas, I el número de inserciones realizadas, y N es el número total de palabras/carácteres en el *ground truth*.

$*$ **NER:** Para NER se usan métricas clásicas como precision, recall, y F1-Score.[60] Sin embargo, una evaluación basada únicamente en coincidencias exactas puede ser demasiado estricta, pues no distingue entre diferentes tipos de error -por ejemplo, cuando el modelo detecta correctamente los límites de la entidad pero falla en asignar la categoría adecuada- Debido a esto, siguiendo el esquema de evaluación introducido en SemEval 2013,[61] los resultados se reportan bajo cuatro escenarios distintos:

- **Strict:** Coincidencia exacta en los límites de la palabra y el tipo de la entidad.
- **Exact:** Coincidencia exacta en los límites de la palabra sin importar el tipo de entidad.
- **Partial:** Coincidencia parcial de límites, sin importar el tipo.

---

59 Rista y Kadriu, ver n. 20.

60 Keraghel, Morbieu y Nadif, ver n. 12.

61 Suresh Manandhar y Deniz Yuret, eds. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, jun. de 2013.

- **Type:** Cierta superposición entre la entidad etiquetada por el sistema y el *ground truth*.

Para calcular las métricas según este sistema se definen 5 conceptos:

- Correcto (COR): Ambos son iguales;
- Incorrecto (INC): La salida del sistema y el *ground truth* no coinciden;
- Parcial (PAR): El sistema y el *ground truth* son algo similares, pero no iguales;
- Faltante (MIS): El sistema no captura una etiqueta del *ground truth*.
- Espurio (SPU): El sistema produce una respuesta que no existe en el *ground truth*.

Luego se calcula $POS = COR + INC + PAR + MIS$, que corresponde al número de entidades *ground truth* y $ACT = COR + INC + PAR + SPU$ que representa el número de entidades predichas. Con esto, obtenemos las siguientes métricas:

$$\text{Precision} = \frac{COR + 0{,}5 \cdot PAR}{ACT}, \tag{6}$$

$$\text{Recall} = \frac{COR + 0{,}5 \cdot PAR}{POS}, \tag{7}$$

$$\text{F1} = \frac{2PR}{P + R}. \tag{8}$$

$*$ **Censurado:** Para el caso de la anonimización también se utilizaron precision, recall y F1 pero adaptadas a la tarea de Censura. En este contexto, un *True Positive* (TP) se considera cuando una palabra sensible (presente como entidad en el *ground truth)* ha sido correctamente censurada por el modelo, es decir, no aparece en el texto resultante. Un *False Negative* (FN) ocurre cuando una entidad sensible no fue

censurada y sigue visible, un *False Positive* (FP) se define como cualquier censura adicional que no corresponde a una entidad anotada, independientemente de su tipo o contenido, es decir, si el modelo censura más entidades de las necesarias, se penaliza con falsos positivos. Esta evaluación no considera si la entidad fue correctamente clasificada ni si el span coincide exactamente con el original en la ubicación, por lo que es más laxa que las métricas tradicionales en tareas de reconocimiento de entidades. El objetivo principal es verificar si la información sensible fue eliminada del texto, más allá de la forma exacta en que se anotó.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{10}$$

$$\text{F1} = \frac{2PR}{P + R}. \tag{11}$$

4.3. ESTUDIOS DE ABLACIÓN

Para evaluar el aporte de distintas decisiones de diseño del modelo se hicieron varios estudios de ablación durante el entrenamiento: (1) sustituir la función de pérdida *cross-entropy* por *focal loss* con el fin de manejar el desbalance de clases —en nuestro caso, entre tokens de entidad y no-entidad—; (2) entrenar el modelo sobre datos ya censurados, de modo que la predicción se limite a la categoría de la entidad sin incluir el texto correspondiente.

Entrenando NER con Focal Loss: Durante los experimentos iniciales observamos que el modelo entrenado con *cross entropy* presentaba dos limitaciones: (1) una alta tasa de errores al asignar las etiquetas de entidad correctas, y (2) una baja

capacidad para predecir tokens de entidad, es decir, tendía a ignorarlos frente a los tokens normales o de contexto. Para mitigar este desbalance, se empleó la *focal loss* bajo la hipótesis de que, frente al desbalance de clases entre distintas categorías de entidad como entre tokens que no son de entidad, le asignará más peso a estos ejemplos difíciles. La focal loss se entrenó con hyperparámetros $\gamma = 2$ y $\alpha = 0{,}25$.

Entrenando con censura: Entrenamos el modelo tanto con *cross entropy* como con *focal loss* para que aprendiera a censurar directamente el texto correspondiente a las entidades. Esta estrategia resulta preferible como alternativa a aplicar un post-procesamiento externo de censura, ya que la supresión de información sensible se incorpora en la propia predicción del modelo, evitando exponer o almacenar el contenido original de las entidades.

Los resultados de nuestros experimentos de ablación se presentan en el Cuadro 3 para la tarea de ASR, en el Cuadro 4 para NER, y en el Cuadro 5 para NER con censura.

Modelo	**WER**	**CER**
Cross-entropy + NER	$8{,}30$	$4{,}55$
Focal loss + NER	$\mathbf{7{,}60}$	$\mathbf{3{,}88}$
Cross entropy + censura	$8{,}55$	$6{,}38$
Focal loss + censura	$11{,}63$	$8{,}57$

Cuadro 3. Resultados del estudio de ablación para ASR. Se comparan cinco configuraciones: Cross entropy con identificación de entidades, Focal loss con identificación de entidades, Cross entropy con censura incluida y Focal loss con censura incluida. Se muestran el WER y el CER para cada configuración.

Analizando los resultados, se observa que el uso de *focal loss* mejora la transcripción automática, reduciendo WER y CER en casi 1 % respecto a *cross-entropy*, lo que indica una mayor precisión general en el ASR. Sin embargo, la aplicación de censura degrada significativamente el rendimiento, aumentando los errores en la transcripción. En el Cuadro 4 se observa que, aunque la *focal loss* ofrece beneficios para ASR,

Modelo	Modo	Precision	Recall	F1
Cross entropy	strict	**55,58**	**55,17**	**54,70**
Cross entropy	exact	**61,86**	**61,51**	**60,81**
Cross entropy	partial	**63,40**	**63,26**	**62,38**
Cross entropy	type	**57,69**	**57,56**	**56,87**
Focal loss	strict	53,16	54,28	53,15
Focal loss	exact	58,55	60,22	58,67
Focal loss	partial	60,20	62,10	60,35
Focal loss	type	55,36	56,74	55,37

Cuadro 4. Resultados del estudio de ablación para NER. Se comparan dos configuraciones: Cross entropy y Focal loss. Se muestran Precision, Recall y F1 para los 4 modos de evaluación planteados: *strict, exact, partial y type*.

Modelo	Precision	Recall	F1
Focal loss	$62,02$	$\mathbf{85,32}$	$71,83$
Cross entropy	$\mathbf{71,24}$	$81,68$	$\mathbf{76,10}$

Cuadro 5. Resultados del estudio de ablación para censura. Se comparan tres configuraciones para censura: Cross entropy y Focal loss. Se muestran Precision, Recall y F1. No se evalúa el acierto ni la ubicación de las entidades sino la eliminación de información sensible.

existe un trade-off en la tarea de reconocimiento de entidades (NER): frente a *cross-entropy*, los resultados son competitivos pero ligeramente inferiores, *cross-entropy* supera consistentemente a *Focal Loss* en las cuatro métricas de evaluación. Para subrayar, el modo que tuvo un mejor rendimiento en ambos modelos fue el *partial*; esto indica que el modelo logra identificar entidades parcialmente, aunque no siempre coincida exactamente con la anotación esperada. Sirva de ejemplo la entidad 'Mariana Afanador', si el modelo predice únicamente 'Mariana' como entidad esta predicción sería considerada un acierto en el modo *partial*, pero no en los modos *strict* o *exact*, los cuales exigen coincidencia completa en el texto de la entidad. Esta diferencia entre modos ejemplifica la sensibilidad de las métricas con errores de segmentación o etiquetado incompleto, comunes en NER. En consecuencia, el uso de los múltiples modos de evaluación permite obtener una visión más variada del comportamiento

del modelo, diferenciando entre errores críticos y errores más tolerables.

El Cuadro 5 muestra los resultados del estudio de ablación sobre los modelos de censura. Bajo nuestra métrica, ambos modelos obtienen un desempeño muy sobresaliente, especialmente en *recall*, donde logran eliminar la mayoría de las entidades sensibles. No obstante, se observa que ambos sacrifican *precision* en favor de un mayor *recall*, lo que sugiere que tienden a censurar más de la cuenta, posiblemente por el tipo de entrenamiento: al entrenar con transcripciones ya censuradas (por ejemplo<person>> en lugar de <person>Miguel<person>>), el modelo pierde parte del contexto léxico. Esto puede llevarlo a 'suponer' qué entidades deben censurarse, en lugar de apoyarse en la información real, aumentando así la probabilidad de falsos positivos.

Aunque la métrica utilizada es laxa —ya que sólo evalúa si las entidades censuradas aparecen en el texto final y no su clasificación o posición exacta—, los resultados sugieren que los modelos cumplen razonablemente bien su objetivo principal: eliminar información sensible del texto, destaca especialmente el modelo *cross-entropy* con un F1 de 76,10 %.

4.3.1. Resultados cuantitativos Tras evaluar el rendimiento de nuestro modelo en el *benchmark* propio, procedimos a compararlo con otros modelos del estado del arte. Específicamente, comparamos el rendimiento de WhisperNER estándar y Whisper con un módulo de post-procesamiento para NER. El Cuadro 6 presenta los resultados de estas pruebas para ASR y el Cuadro 7 para NER.

Modelo	**WER**	**CER**
Whisper *(large-v2)*	$8{,}57$	$5{,}75$
WhisperNER	$18{,}37$	$13{,}76$
Nuestro *(Focal NER)*	$\mathbf{7{,}60}$	$\mathbf{3{,}88}$

Cuadro 6. Comparación del rendimiento del modelo contra WhisperNER y Whisper v2 large. Se muestran los resultados de *WER*, *CER*.

Modelo	Modo	Precision	Recall	F1
NER *(flair-spanish-large)*	strict	**74,79**	**73,18**	**73,36**
NER *(flair-spanish-large)*	exact	**76,22**	**74,61**	**74,77**
NER *(flair-spanish-large)*	partial	**77,02**	**75,42**	**75,55**
NER *(flair-spanish-large)*	type	**76,11**	**74,53**	**74,66**
NER *(GLiNER multi-v2.1)*	strict	50,67	42,70	44,75
NER *(GLiNER multi-v2.1)*	exact	53,73	44,99	47,21
NER *(GLiNER multi-v2.1)*	partial	55,17	46,08	48,34
NER *(GLiNER multi-v2.1)*	type	52,57	44,19	46,26
WhisperNER	strict	27,93	27,35	25,96
WhisperNER	exact	30,26	29,76	27,95
WhisperNER	partial	31,38	31,18	29,08
WhisperNER	type	29,45	29,28	27,49
Nuestro *(cross-entropy NER)*	strict	55,58	55,17	54,70
Nuestro *(cross-entropy NER)*	exact	61,86	61,51	60,81
Nuestro *(cross-entropy NER)*	partial	63,40	63,26	62,38
Nuestro *(cross-entropy NER)*	type	57,69	57,56	56,87

Cuadro 7. Comparación del rendimiento del modelo contra WhisperNER, y modelos NER post-procesamiento *flair ner spanish large*[62]y *GLiNER*.[63] Se muestran Precision, Recall y F1 para los 4 modos de evaluación planteados: *strict*, *exact*, *partial* y *type*.

Analizando los resultados de los Cuadros 6 y 7, se observa que nuestro modelo alcanza un desempeño superior en la tarea de ASR frente a los modelos de referencia. En particular, el modelo con *Focal NER* logra un WER de $7{,}60\,\%$ y un CER de $3{,}88\,\%$, superando tanto a Whisper *large-v2* como a WhisperNER en este benchmark. Esto evidencia que la estrategia de entrenamiento y la configuración empleada permiten un reconocimiento de voz altamente preciso.

En la tarea de NER, nuestro modelo demuestra un rendimiento competitivo, a pesar de no superar a modelos preentrenados en grandes corpus de español como

---

[63] Stefan Schweter y Alan Akbik. *FLERT: Document-Level Features for Named Entity Recognition*. 2020. arXiv: `2011.06993 [cs.CL]`

[63] Urchade Zaratiana et al. *GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer*. 2023. arXiv: `2311.08526 [cs.CL]`

flair-spanish-large[64] —ajustado sobre el subset español de CoNLL-03 Spanish,[65] con 11.758 archivos de texto—, logra superar al estado del arte GLINeR multilenguaje 2.1[66] en los cuatro modos de evaluación (*strict, exact, partial y type*). Es importante destacar que nuestro enfoque "todo en uno" mantiene resultados sólidos y consistentes habiendo sido entrenados 8.028.160 parámetros de un total de 1.551.169.280, logrando un rendimiento competente aún con recursos limitados; y además, esta diferencia de rendimiento frente a *Flair* puede atribuirse en gran medida a la escasez de datos con entidades utilizados para ajustar nuestro modelo (véase la Figura 9(a) y el Cuadro 2) en contraste con el entrenamiento masivo sobre texto que recibió *Flair*. Para suplementar el análisis de nuestro modelo, se hizo una comparación de resultados para todas las regiones en el Cuadro 8, para cada etiqueta de entidad en el Cuadro 9, y con otros modelos del estado del arte contra un dataset fuera del conjunto de datos recopilado: FLEURS[67] en su partición de test, español España en el Cuadro 10. Las métricas de los demás modelos fueron tomadas por la comunidad Open ASR Leaderboard.[68]

Primero que todo, el modelo presentó su mejor rendimiento en la región de Tunja, lo cual podría explicarse por el reducido número de segmentos evaluados. En contraste, Bogotá y Bucaramanga mostraron los peores desempeños, alcanzando un WER del

---

64 Schweter y Akbik, *FLERT: Document-Level Features for Named Entity Recognition*, ver n. 63.

65 Erik F. Tjong Kim Sang. "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition". En: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 2002.

66 Zaratiana et al., ver n. 63.

67 Alexis Conneau et al. "FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech". En: *arXiv preprint arXiv:2205.12446* (2022).

68 Vaibhav Srivastav et al. *Open Automatic Speech Recognition Leaderboard*. `https://huggingface.co/spaces/hf-audio/open_asr_leaderboard`. 2023.

Ciudad/Región	WER	CER	# Audios	Duración (min)
Arauca	6,37	2,32	30	10,06
Barranquilla	6,67	2,97	100	35,63
Bogotá	12,33	8,02	98	36,30
Bucaramanga	9,56	4,40	83	30,36
Cali	8,03	3,84	114	42,11
Cúcuta	6,03	2,99	94	33,29
Medellín	6,52	3,34	93	31,28
Neiva	7,40	3,34	51	17,60
Pasto	7,13	3,86	53	18,30
Quibdó	6,66	3,76	57	20,44
Tunja	**4,19**	**1,72**	35	12,14
Yopal	6,80	3,00	48	16,81
San Andrés	4,97	2,60	40	15,07

Cuadro 8. Rendimiento del modelo *Focal NER* para ASR por región, mostrando también el número de segmentos evaluados y duración total.

Etiqueta de entidad	Total entidades	Precision	Recall	F1
PER	327	36,88	36,36	35,93
ORG	226	18,87	15,88	16,60
LOC	985	**49,00**	**53,02**	**49,79**
MISC	422	19,90	20,02	19,40

Cuadro 9. Rendimiento del modelo *Cross NER* para NER *Strict* por tipo de entidad, mostrando también el número de entidades totales evaluadas.

Modelo	WER
Whisper *(large-v3)*	**2,62**
Voxtral Mini *(3VV-2507)*	3,34
ElevenLabs Scribe *(v1)*	7,65
Nuestro *(focal NER)*	8,86

Cuadro 10. Comparación del rendimiento ASR sobre el dataset FLEURS (test-es).

12,33 % y 9,56 % respectivamente. Este resultado puede atribuirse a dos factores principales luego de una revisión del conjunto de datos en su partición de test: (1) la presencia de varios audios con calidad deficiente en estas dos regiones, lo que dificulta la transcripción automática; y (2) inconsistencias detectadas en las

transcripciones de referencia —por ejemplo, omisiones o errores de escritura— que elevan las métricas de error. Esta última observación destaca la sensibilidad de las métricas automáticas como WER frente a la calidad del corpus de referencia, especialmente en evaluaciones fuera de corpus estandarizados. Al revisar el Cuadro 9, se observa que la etiqueta **"location"** (LOC) alcanza el mayor F1-score, con un 49,79 %, seguida de **"person"** (PER), con un 35,93 %. Estos resultados son coherentes con el desbalance de clases presente en el conjunto de datos. Por otro lado, en el Cuadro 10 los resultados muestran que nuestro modelo Focal NER con WER = 8,86 % no supera los niveles de los modelos de punta reportados (Whisper large, Voxtral, ElevenLabs) en el conjunto de prueba en español de FLEURS. Cabe destacar que estos modelos fueron preentrenados con cientos de miles de horas de audio multilingüe, lo que les permite capturar patrones acústicos y lingüísticos con un alto grado de generalización. En cambio, nuestro modelo, sin acceso a tales cantidades de datos y sin preentrenamiento específico en FLEURS, logra resultados robustos, lo cual resalta la eficacia del enfoque propuesto.

4.3.2. Resultados cualitativos

Ejemplo de la pipeline: Para ilustrar de forma clara y visual el funcionamiento completo del sistema, primero se presenta a continuación una comparación paso a paso del *pipeline* de anonimización en la Figura 10. Se muestran cuatro etapas fundamentales: (1) el texto original o *ground truth*, (2) la transcripción automática del audio con anotación de entidades (NER), (3) el resultado del post-procesamiento para censura sobre la transcripción, y (4) la salida directa del modelo que integra transcripción y censura. Esta comparación permite visualizar el flujo completo desde el audio hasta la generación del texto anonimizado, así como observar fortalezas y errores en cada etapa. Para mayor claridad, cada etapa se presenta en un bloque separado y con colores diferenciados.

Ground Truth (Texto Original)

LOC
Caño Gualabao , lo más emblemático de este hermoso pueblo. Monumento PER
de Inocencio Chincá , uno de los libertadores de la batalla de LOC Boyacá . Estatua hace homenaje a las mujeres indígenas y MISC Giraras en el municipio de LOC Tame . LOC Parque Central , pero como está en remodelación, tenemos la esquina de LOC Cayo Mario .

Transcripción con Cross NER

LOC
Taño Gualabao , lo más emblemático de este hermoso pueblo. LOC
Monumento de Inocencio Chincá , uno de los libertadores de la MISC Batalla Boyacá . LOC Estaco , acá se homenajea a las mujeres indígenas y giraras en el municipio de LOC Tame . LOC Parque Central , pero como está en remodelación, tenemos la esquina de LOC Callomario .

Censura post-procesamiento

LOC
<location> , lo más emblemático de este hermoso pueblo. LOC <location> , uno de los libertadores de la MISC <miscellaneous> . LOC <location> , acá se homenajea a las mujeres indígenas y giraras en el municipio de LOC <location> . LOC <location> , pero como está en remodelación, tenemos la esquina de LOC <location> .

Modelo con censura integrada

LOC
<loc> , lo más emblemático de este hermoso pueblo. LOC
<loc> , uno de los libertadores de la batalla de LOC <loc> . LOC
<loc> homenaje a las mujeres indígenas y giraras en el municipio de LOC <loc> . LOC
<loc> , pero como está en remodelación tenemos la esquina de LOC <loc> .

Figura 10. Flujo completo desde la transcripción hasta la anonimización final. Se muestra el texto original, la transcripción con anotación de entidades, el resultado del post-procesamiento de censura, y la salida directa del modelo con censura integrada.

Ejemplos cualitativos: Para complementar nuestro análisis cuantitativo, presentamos una evaluación cualitativa de los resultados obtenidos por nuestro método Focal NER, modelo Censura, y de la anotación temporal precisa con diarización de hablantes en algunos audios de la partición test. Se seleccionaron cinco audios en total. Los tres primeros pertenecen a evaluaciones del modelo Focal NER para evaluar ASR-NER, mientras que los dos restantes fueron seleccionados con el objetivo de analizar ASR y el comportamiento del modelo de censura, es decir, cómo y cuándo decide eliminar información sensible. En los tres primeros casos se emplearon muestras en escenarios menos controlados (ruido, voces superpuestas, etc.), mientras que en los dos últimos se utilizaron muestras con múltiples entidades, con el fin de examinar la efectividad del proceso de anonimización.

Los resultados fueron divididos en 3 tablas: la primera presenta las transcripciones resultantes divididas en segmentos de cada audio, la segunda muestra los timestamps y hablantes asignados por segmento de cada audio y la tercera es el Cuadro 11, que reporta las métricas de cada ejemplo. La notación usada es la siguiente: Para la tarea de transcripción se indican en rojo los caracteres o palabras añadidos incorrectamente por el modelo, en azul aquellos que fueron omitidos, y en negro todo lo correcto. En la identificación de entidades, cada tipo se marcó con un color: **“person”** se denota con el color magenta, **“location”** con el color violeta, **“organization”** con verde y **“miscellaneous”** con naranja. Además, se añadieron 2 indicadores adicionales para representar errores tanto en NER como en censura: *wrong*, en rojo, se utiliza cuando el modelo marca incorrectamente una palabra como entidad o censura una palabra que no debía; y *missing*, en azul, indica la ausencia de una etiqueta de entidad o una omisión en la censura de una palabra sensible. En la diarización, los hablantes incorrectos aparecen en rojo y los correctos en verde.

Por último, en los ejemplos cualitativos del modelo con censura, se añadió manualmente el contenido original censurado dentro de las etiquetas con fines ilustrativos;

originalmente, el modelo devuelve las etiquetas de censura sin revelar el texto sensible. La inclusión del texto original en los ejemplos es únicamente para facilitar la interpretación y evaluar visualmente si la censura fue aplicada correctamente.

Segmento 0: Es el impresionante restaurante mexicano en LOC Bucaramanga que debes conocer ahora mismo porque realmente te transporta a LOC Ciudad de México .

Segmento 1: Allí manejan los mejores tacos de birria de todo el LOC área metropolitana , sin contar su nuevo y exquisito producto, la increíble pizza de birria.

Segmento 2: Perfecta recomendación para invitar a tu pareja, amigos, familia o a la abuelita MISSING cleotilde .

Segmento 3: Ellos son la MISSING taquería chula , papá.

Segmento 4: Ubicados específicamente en un LOC parqueadero , en la LOC carrera 28 #4968 , LOC Barrio Sotomayor , no hay pérdida, hay mucho espacio para moto.

Figura 11. Audio 1 Focal NER: Resultados para ASR con NER

Segmento	Hablante	Timestamp(s)	Transcripción
0	SPEAKER_01	0.03 – 6.45	Es el . . .
1	SPEAKER_01	6.47 – 14.07	Allí manejan . . .
2	SPEAKER_01	14.11 – 18.32	Perfecta . . .
3	SPEAKER_01	18.34 – 20.72	Ellos son . . .
4	SPEAKER_01	20.74 – 28.82	Ubicados . . .

Figura 12. Audio 1 Focal NER: Resultados para timestamps y diarización

En audios claros y fluidos con múltiples entidades, como el Audio 1, el modelo realiza una transcripción casi perfecta, con errores mínimos y pocas omisiones, como se muestra en el Cuadro 11 (WER $0,256$, CER $0,080$). A pesar de ello, las métricas de NER muestran resultados moderados (F1-Score $0,33$), ello debido a entidades sin etiquetar y entidades etiquetadas erróneamente, por ejemplo se etiquetó “carrera 28” como *location*, sin embargo en el *ground truth* es una etiqueta *miscellaneous*.

Segmento 0: Les saludamos hoy desde la hermosa e inigualable LOC Barranquilla .

Segmento 1: Les contamos que vinimos a carnavales, estamos acá en el malecón del

.¿tiburón?

Segmento 2: .

Segmento 3: .

Segmento 4: ¿figura?

Segmento 5: ¿como?

Segmento 6: En el malecón del LOC Caimán .

Segmento 7: Y hay un evento de por allá que es el de LOC Colombia , bajen por allá WRONG hombres .

Segmento 8: Bueno, como ustedes ya saben y si no saben, el río MISSING Magdalena , creo que se acaba en, sí, se acaba en MISSING Barranquilla .

Figura 13. Audio 2 Focal NER: Resultados para ASR con NER

En el Audio 2, que presenta ruido e interferencias significativas donde incluso hasta

la comprensión humana es limitada y varios segmentos no fueron transcritos en el *ground truth*, el modelo logra reconocer palabras y entidades que a veces se escapan a un oyente humano (WER $0,393$, CER $0,214$), aunque con menor precisión en NER (F1-Score $0,17$). Esto evidencia que incluir audios variados (incluso de mala calidad) durante el entrenamiento es crucial para que el modelo generalice y se desempeñe bien en distintos escenarios.

Segmento	Hablante	Timestamp (s)	Transcripción
0	SPEAKER_00	0.03 – 6.22	Les . . .
1	SPEAKER_00	6.24 – 10.63	Les . . .
2	UNKNOWN	10.65 – 10.93	·
3	SPEAKER_00	10.95 – 11.01	·
4	SPEAKER_00	11.05 – 11.45	¿figura?
5	SPEAKER_00	11.79 – 12.67	¿como?
6	SPEAKER_00	12.69 – 14.97	En el . . .
7	SPEAKER_00	15.43 – 21.14	Y hay . . .
8	SPEAKER_00	21.16 – 29.84	Bueno, . . .

Figura 14. Audio 2 Focal NER: Resultados para timestamps y diarización

Algo a destacar es que la diarización tiende a fallar cuando los hablantes se solapan, una característica presente en este audio, complicando la diferenciación entre voces simultáneas. De manera similar, en segmentos con silencio (como los segmentos 2 y 3), el modelo tiende a “rellenar” estos espacios, generando puntos u otros símbolos. Este comportamiento es típico de su arquitectura auto-regresiva, tendiendo a producir

"halucinaciones" ante ausencia de señal de audio.[69]

Segmento 0: ¿Qué tal Familia Cumbiera?

Segmento 1: Bueno, esto se llama Pin Pong. con. una cumbiera más.

Segmento 2: ¿Nombre completo?.

Segmento 3: Geraldine Paola Diaz Gamarra (PER) .

Segmento 4: ¿De Aquí (WRONG) de Bucaramanga (LOC) ?

Segmento 5: Sí, a mucho honor Bumanguesa (PER) .

Segmento 6: Edad.

Segmento 7: 22 años.

Segmento 8: Uy, esta jovencitica ¿no Geraldine (MISSING) ?

Segmento 9: Sí.

Segmento 10: ¿Profesión?.

Figura 15. Audio 3 Focal NER: Resultados para ASR con NER

En situaciones más desafiantes, como en el Audio 3 —que incluye un acento marcado y varios hablantes— se incrementan los errores (WER $0,475$, CER $0,112$) palabras omitidas o escritas de forma distinta (como en el caso de Pon, en vez de Pong, símbolos de interrogación faltantes, etc). Sin embargo, el modelo aún identifica correctamente varias entidades y mantiene coherencia en la transcripción (F1-Score $0,50$).

---

[69] Mateusz Barański et al. "Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio". En: (ene. de 2025). DOI: 10.48550/arXiv.2501.11378.

Segmento	Hablante	Timestamp(s)	Transcripción
0	SPEAKER_00	0.03 – 6.48	¿Qué tal . . .
1	SPEAKER_00	6.54 – 10.73	Bueno, esto . . .
1a	SPEAKER_00	–	Pon
1b	SPEAKER_00	–	Una cumbiera . . .
2	SPEAKER_00	11.59 – 12.27	Nombre . . .
3	SPEAKER_01	12.29 – 14.45	Geraldine . . .
4	SPEAKER_01	14.47 – 16.58	¿De aquí de . . .
5	SPEAKER_01	16.60 – 19.22	Sí a mucho . . .
6	SPEAKER_01	19.24 – 19.34	Edad.
7	SPEAKER_01	19.36 – 19.66	22 años.
8	SPEAKER_01	19.68 – 22.38	Uy esta . . .
9	SPEAKER_01	22.40 – 22.62	Sí.
10	SPEAKER_00	23.73 – 24.27	Profesión.

Figura 16. Audio 3 Focal NER: Resultados para timestamps y diarización

Similar al Audio 2, múltiples hablantes que se solapan suelen clasificarse erróneamente. Una segunda posible causa a los errores son los datos sobre los que fue entrenado el modelo de diarización, los cuales eran en inglés y pueden afectar su rendimiento en otros idiomas como el español.

Los dos audios restantes se evaluaron con nuestras métricas de censura para observar si el modelo elimina efectivamente las entidades sensibles.

Segmento 0: Cabe destacar que los fuertes rumores que andan en las redes sociales es que ni el señor [Medellín]$^{\text{MISC}}$ ni [Lupita tiktok]$^{\text{MISC}}$ supuestamente ya pueden darle declaraciones a otros medios de comunicación porque ellos vendieron la exclusiva a [Chavana]$^{\text{MISC}}$.

Segmento 1: Esto no está confirmado pero lo que sí es que todas las noches ellos están conectando con [Chavana]$^{\text{MISC}}$ para darle actualización del estado de salud de la pequeña.

Figura 17. Audio 4 Modelo Censura: Resultados para ASR con NER

Segmento	Hablante	Timestamp(s)	Transcripción
0	SPEAKER_00	0.03 – 13.29	Cabe . . . .
1	SPEAKER_00	13.93 – 22.04	Esto . . . .

Figura 18. Audio 4 Modelo Censura: Resultados para timestamps y diarización

Para el Audio 4 se mantienen un WER y CER bajos (WER $0,031$, CER $0,030$), un resultado esperado pues es una muestra fácil en cuanto a transcripción se refiere. Así mismo, se tiene un F1 de 1, es decir se censuraron todas las entidades relevantes. Si evaluamos cualitativamente este ejemplo podemos observar que, a pesar de anonimizar toda la información sensible, a veces no asigna correctamente el tipo de entidad (véase Lupita tiktok como entidad Miscelánea en vez de Persona).

Segmento 0: tu marca y su imagen en todo el mundo y más en un país como [España]$^{\text{LOC}}$ donde está lleno de [colombianos]$^{\text{MISC}}$.

Segmento 1: Esta gente lanza el producto aprovechándose de todos esos [colombianos]$^{\text{MISC}}$ que están allá porque para los que no lo sepan [España]$^{\text{LOC}}$ es el segundo país con mayor cantidad de migrantes [colombianos]$^{\text{MISC}}$ que no pueden venir a [Colombia]$^{\text{LOC}}$ y que les encantaría poder probar los productos.

Segmento 2: No creo tampoco que sea una estrategia comercial de la marca para llegar allá porque estoy seguro que [Frisby]$^{\text{MISC}}$ no lo necesita y no se presta para ese tipo de cosas.

Figura 19. Audio 5 Modelo Censura: Resultados para ASR con NER

Segmento	Hablante	Timestamp(s)	Transcripción
0	SPEAKER_00	0.03 – 6.22	tu marca . . .
1	SPEAKER_00	6.40 – 18.65	Esta gente . . .
2	SPEAKER_00	18.93 – 25.50	No creo . . .

Figura 20. Audio 5 Modelo Censura: Resultados para timestamps y diarización

En el Audio 5 también se tienen un WER y CER bajos (WER $0,022$, CER $0,030$). Similar al Audio 4, se obtiene un F1 de 1, censurando todas las entidades relevantes. Para el caso de este ejemplo se asignan más etiquetas de entidad correctamente (España como *(*location), colombianos como *miscellaneous*)

Audio	WER	CER	F1-Score (Exact) & Censor
Audio 1	0.256	0.080	0.33
Audio 2	0.393	0.214	0.17
Audio 3	0.475	0.112	0.50
Audio 4	0.031	0.030	1.0
Audio 5	0.022	0.030	1.0

Cuadro 11. Métricas de los modelos Focal NER y censura para los cinco audios de evaluación cualitativa, incluyendo WER, CER y F1-Score en reconocimiento de entidades (*exact*) y censura.

4.3.3. Resultados experimentales cualitativos Finalmente, para evaluar la capacidad de generalización de nuestro modelo en escenarios del mundo real, se realizó una validación cualitativa de la tarea de ASR-NER y censura utilizando audios capturados por terceros fuera del conjunto de datos. Esta prueba tiene como objetivo demostrar la robustez y aplicabilidad del método propuesto en entornos no controlados. Se grabaron 6 audios utilizando distintos tipos de micrófonos y con diversos acentos (no solo colombianos). Estas pruebas se hicieron sobre los modelos Focal NER, Cross NER y el modelo de censura.

> Hagamos lo que dije, que cada uno tenga un instintivo tipo, si entro tal, los chicos van a ver cuando se lo den igual. Encima no va a entrar todas las frases en las bolsitas porque son re chicas.

Figura 21. Audio 1 Experimental Focal NER: Resultados para acento argentino mujer

PER
Hola Gerardo , ¿qué tal? Buenos días. Por favor informar a los PER chicos que no estén dando todavía la evaluación que está en de la práctica calificada 1, ¿ya? Voy a crear otra porque al parecer la fecha me ha ganado, ¿ya? Y ya les voy a estar mandando el anuncio. Igual te voy a reforzar contigo para que les digas, ¿ya? Igual compartan compártale este audio con ellos para que entiendan. O comprendan que aún no entren a dar la práctica calificada a WRONG uno , que está ya, este, digamos, vista, ¿no? Esa no la van a dar porque no hay preguntas. ¿Ya? ¡Gracias!

Figura 22. Audio 2 Experimental Focal NER: Resultados para acento peruano hombre

Entonces sería viernes en la tarde tipo 5 pm en la LOC Uis , esto partimos la torta en la LOC Uis , esto una torta chiquitica y luego iríamos a la MISSING 25 y tomamos y pues si, y eso le gastamos al corto gordo. Si, voy a mandar este audio como a todo el mundo, confirmame.

Figura 23. Audio 3 Experimental Cross NER: Resultados para acento bumangués mujer

Hola, buenas noches. Estoy intentando hablar con acento que no sea colombiano (MISC) . Aunque creo que yo no tengo acento colombiano (MISC) . Sí, creo que no. Pero igual es medio parecido. Ay, no sé si lo tenía que hablar en inglés o en español. Ay, pues si me dices que no sea colombiano (MISC) , obvio es en español. Bueno. Eso. Espero te sirva mi audio. Muack.

Figura 24. Audio 4 Experimental Cross NER: Resultados para acento venezolano mujer

Mi nombre es Mariana (PER) , mi número de celular es [NÚMERO DE CELULAR] (MISSING) . Estoy estudiando biología en la Universidad de [ESTADO] (LOC) . Estoy en mi último semestre. Vivo en un conjunto de apartamentos que se llama [NOMBRE DE APARTAMENTOS] (LOC) . Edificio [NÚMERO EDIFICIO] (MISSING) , apartamento [NÚMERO APTO] (MISSING) .

Figura 25. Audio 5 Experimental Censura: Resultados para acento bumangués mujer

Bueno PER [Andrea] , mi nombre es PER [Juan Esteban] , con cédula MISSING [NÚMERO DE CÉDULA] , y bueno, pues, ¿qué te cuento? Voy a hablarte de la serie que estoy viendo porque es lo único que tengo en la cabeza en este momento. ¿Qué te cuento WRONG [TÉRMINO COLOQUIAL] ? Está muy buena esa serie. PER [Andrea] , de verdad, deberías darle un try a la serie porque está muy buena. Eh, no sé, se llama El Chacal, bueno en español se llama El Chacal, en inglés es The Day of Chacal. Y pues nada, ¿qué te digo? Está interesante, o sea tiene una trama chévere, pero siento que es una trama que solo va a durar una temporada, o sea no le van a sacar más temporadas, y lástima porque pues ese cast tiene bastante potencial para otra temporada al menos. Y además fueron 10 capítulos, o sea, me la comí. En dos días. Osea, fue en 10 capítulos, de 40 a 50 minutos. La verdad, muy buena, recomendada, está en MISC [Disney Plus] , por si quieren, les presto mi cuenta. Y ya, eh, espero que te sirva este audio, si necesitas más audio, me avisas. ¡Gracias!

Figura 26. Audio 6 Experimental censura: Resultados para acento bumangués hombre

Como se puede observar, nuestro modelo focal NER demuestra una notable capacidad para generalizar a audios no escuchados previamente e incluso de distintos países de habla hispana, lo que indica una buena capacidad *zero-shot*. En los ejemplos presentados, el modelo rara vez comete errores evidentes, y sus predicciones de entidades resultan coherentes con el contexto, incluso en situaciones informales o con ruido lingüístico (muletillas, coloquialismos). En el caso del modelo con censura, si bien logró anonimizar correctamente algunas entidades, también se observan

varias que no fueron eliminadas, como el número de cédula, número de celular, número de edificio y número de apartamento. Esta omisión es una limitación relevante, aunque esperable, ya que al revisar el conjunto de datos se evidencia que este tipo de entidades no estaban representadas o aparecían muy escasamente. La falta de ejemplos relacionados con esta información durante el entrenamiento probablemente impidió que el modelo aprendiera que debe censurarlas.

También se identifican otros patrones interesantes de comportamiento para los tres modelos: cuando se le suministran *prompts* que contienen entidades que no están presentes en el audio —por ejemplo, *location* en un audio donde no se menciona ningún lugar— el modelo intenta 'forzar' entidades en la salida aunque no correspondan al contenido real. Por otro lado, si no se le indica que debe buscar una entidad específica en el *prompt*, el modelo no genera etiquetas en absoluto, produciendo una salida sin anotaciones. En ambos casos, este comportamiento subraya la necesidad de un diseño cuidadoso del *prompt* para evitar tanto omisiones como alucinaciones de entidades.

Cabe aclarar que todas las entidades detectadas mediante NER en los ejemplos son censuradas posteriormente. No obstante, se presenta el texto original junto con la salida del modelo para permitir una evaluación más completa y transparente de los resultados.

5. CONCLUSIONES

En este trabajo, el desarrollo de un conjunto de datos propio de audio en español colombiano resultó clave para obtener un modelo de ASR adaptado a nuestra realidad lingüística, con diversidad de hablantes, acentos y expresiones locales; además de obtener resultados competitivos para reconocimiento de entidades que llegan a superar algunos métodos del estado del arte. Los experimentos muestran que la elección de diferentes funciones de pérdida, como *cross-entropy* y *focal loss*, tiene un impacto directo en el rendimiento del modelo, afectando de manera distinta la transcripción y el reconocimiento de entidades. Se logró responder la pregunta de investigación y cumplir con los objetivos planteados, desarrollando un modelo de lenguaje natural capaz de obtener transcripciones anonimizadas, con identificación de entidades y anotación con marcas de tiempo por intervenciones, a partir de grabaciones de audio del español de Colombia. Estos resultados destacan la relevancia de combinar conjuntos de datos específicos con estrategias de *fine-tuning* y pérdida adaptativa para tareas multilingües y multidominio, y sientan las bases para futuras mejoras en modelos que integran ASR junto con otras tareas de NLP. Puesto que distintos modelos mostraban un mejor desempeño según la tarea, se añadió un argumento que permite seleccionar el modelo más adecuado según la prioridad: ASR, NER, censored-NER o post-procesamiento censurado.

6. TRABAJO FUTURO

Como se observó, mejoras en ASR no siempre se traducen en mejoras en NER, y viceversa, un comportamiento previamente reportado.[70] Debido a esto, como trabajo futuro sería interesante explorar funciones de pérdida híbridas que permitan maximizar el rendimiento de ASR y NER simultáneamente sin comprometer ninguna. La incorporación de censura dentro del modelo a veces afecta la detección de entidades; investigar ajustes arquitectónicos o de entrenamiento que preserven el contexto necesario para NER mientras se aplica censura podría mejorar el rendimiento del modelo. También resulta relevante abordar el desbalance de clases y la presencia de etiquetas NER generadas automáticamente, expandiendo y corrigiendo el conjunto de datos base para mejorar la calidad e incluso poder llegar a tener un *benchmark* colombiano de ASR + NER. Finalmente, sería valioso investigar la extensión del modelo a otras tareas de NLP y evaluar su generalización en diferentes dialectos y dominios del español, ampliando así su aplicabilidad y robustez.

[70] Ayache et al., ver n. 30.

BIBLIOGRAFÍA

Ahlawat, Harsh, Naveen Aggarwal y Deepti Gupta. “Automatic Speech Recognition: A survey of deep learning techniques and approaches”. En: *International Journal of Cognitive Computing in Engineering* 6 (2025), págs. 201-237. DOI: https://doi.org/10.1016/j.ijcce.2024.12.007 (vid. págs. 15, 20, 25).

Alharbi, Sadeen et al. “Automatic Speech Recognition: Systematic Literature Review”. En: *IEEE Access* PP (sep. de 2021), págs. 1-1. DOI: 10.1109/ACCESS.2021.3112535 (vid. pág. 14).

Ayache, Gil et al. “WhisperNER: Unified Open Named Entity and Speech Recognition”. En: *arXiv preprint arXiv:2409.08107* (2024) (vid. págs. 27, 28, 31, 33, 40, 72).

Baevski, Alexei et al. “wav2vec 2.0: a framework for self-supervised learning of speech representations”. En: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Vancouver, BC, Canada: Curran Associates Inc., 2020 (vid. pág. 36).

Bain, Max et al. “WhisperX: Time-Accurate Speech Transcription of Long-Form Audio”. En: *INTERSPEECH 2023* (2023) (vid. págs. 27, 29, 33).

Barański, Mateusz et al. “Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio”. En: (ene. de 2025). DOI: 10.48550/arXiv.2501.11378 (vid. pág. 62).

Basak, Sneha et al. “Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems”. En:

Computer Modeling in Engineering Sciences 135 (oct. de 2022), págs. 1-37. DOI: 10.32604/cmes.2022.021755 (vid. pág. 26).

Bredin, Hervé et al. “pyannote.audio: neural building blocks for speaker diarization”. En: *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*. Barcelona, Spain, mayo de 2020 (vid. págs. 33, 36).

Conneau, Alexis et al. “FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech”. En: *arXiv preprint arXiv:2205.12446* (2022) (vid. pág. 54).

yt-dlp Project. *yt-dlp: A feature-rich command-line audio/video downloader*. https://github.com/yt-dlp/yt-dlp. 2025 (vid. pág. 37).

FFmpeg Developers. *FFmpeg: multimedia framework*. https://ffmpeg.org/. 2025 (vid. pág. 37).

Hu, Edward J et al. “LoRA: Low-Rank Adaptation of Large Language Models”. En: *International Conference on Learning Representations*. 2022 (vid. págs. 31, 32, 41).

Hu, Ke et al. “Word Level Timestamp Generation for Automatic Speech Recognition and Translation”. En: *Interspeech 2025*. 2025, págs. 2565-2569. DOI: 10.21437/Interspeech.2025-869 (vid. pág. 26).

— “Word Level Timestamp Generation for Automatic Speech Recognition and Translation”. En: *Interspeech 2025*. 2025, págs. 2565-2569. DOI: 10.21437/Interspeech.2025-869 (vid. pág. 27).

Huang, Zhiqiang, Liang He, Yu Yang et al. “Application of machine reading comprehension techniques for named entity recognition in materials science”. En: *Journal of Cheminformatics* 16.76 (2024). DOI: 10.1186/s13321-024-00874-5 (vid. pág. 42).

Keraghel, Imed, Stanislas Morbieu y Mohamed Nadif. *Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study*. 2024 (vid. págs. 19, 47).

Kumar, Shashi et al. “TokenVerse: Towards Unifying Speech and NLP Tasks via Transducer-based ASR”. En: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, nov. de 2024, págs. 20988-20995. DOI: 10.18653/v1/2024.emnlp-main.1167 (vid. pág. 26).

Malik, Mishaim et al. “Automatic speech recognition: a survey”. En: *Multimedia Tools and Applications* 80 (mar. de 2021), págs. 1-47. DOI: 10.1007/s11042-020-10073-7 (vid. pág. 23).

Manandhar, Suresh y Deniz Yuret, eds. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, jun. de 2013 (vid. pág. 47).

Martinez, Angel Mario Castro y Marc René Schädler. “Why do ASR Systems Despite Neural Nets Still Depend on Robust Features”. En: *Interspeech 2016*. 2016, págs. 1883-1887. DOI: 10.21437/Interspeech.2016-1552 (vid. pág. 23).

Mishchenko, Konstantin y Aaron Defazio. “Prodigy: An Expeditiously Adaptive Parameter-Free Learner”. En: *Proceedings of the 41st International Conference on Machine Learning*. Ed. por Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, 21–27 Jul de 2024, págs. 35779-35804 (vid. pág. 41).

Moskvitch, Katia. *The machines that learned to listen*. https://www.bbc.com/future/article/20170217-the-machines-that-learned-to-listen. Último acceso: 16 de noviembre de 2025. 2017 (vid. pág. 14).

Nakayama, Hiroki et al. *doccano: Text Annotation Tool for Human*. Software available from https://github.com/doccano/doccano. 2018 (vid. pág. 19).

National Academies of Sciences, Engineering y Medicine. *Voice Communication Between Humans and Machines*. Applications of Voice-Processing Technology in Telecommunications. Washington, DC: The National Academies Press, 1994. DOI: 10.17226/2308 (vid. pág. 14).

O'Shaughnessy, Douglas. "Speaker Diarization: A Review of Objectives and Methods". En: *Applied Sciences* 15.4 (2025). DOI: 10.3390/app15042002 (vid. pág. 26).

OpenAI. *Introducing Whisper*. https://openai.com/index/whisper/. Último acceso: 16 de noviembre de 2025. 2022 (vid. pág. 23).

Park, Taejin et al. *Sortformer: A Novel Approach for Permutation-Resolved Speaker Supervision in Speech-to-Text Systems*. 2025. arXiv: 2409.06656 [eess.AS] (vid. pág. 27).

Qin, Libo et al. *Large Language Models Meet NLP: A Survey*. 2024 (vid. págs. 15, 18).

Radford, Alec et al. "Robust Speech Recognition via Large-Scale Weak Supervision". En: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul de 2023, págs. 28492-28518 (vid. págs. 15, 23, 38).

Rista, Amarildo y Arbana Kadriu. “Automatic Speech Recognition: A Comprehensive Survey”. En: *SEEU Review* 15 (dic. de 2020), págs. 86-112. DOI: 10.2478/seeur-2020-0019 (vid. págs. 23, 47).

Schweter, Stefan y Alan Akbik. *FLERT: Document-Level Features for Named Entity Recognition*. 2020. arXiv: 2011.06993 [cs.CL] (vid. pág. 39).

— *FLERT: Document-Level Features for Named Entity Recognition*. 2020. arXiv: 2011.06993 [cs.CL] (vid. págs. 53, 54).

SeleniumHQ. *Selenium WebDriver*. https://www.selenium.dev/. 2025 (vid. pág. 37).

Shafey, Laurent, Hagen Soltau e Izhak Shafran. “Joint Speech Recognition and Speaker Diarization via Sequence Transduction”. En: sep. de 2019, págs. 396-400. DOI: 10.21437/Interspeech.2019-1943 (vid. pág. 27).

Srivastav, Vaibhav et al. *Open Automatic Speech Recognition Leaderboard*. https://huggingface.co/spaces/hf-audio/open_asr_leaderboard. 2023 (vid. pág. 54).

Team, Silero. *Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. https://github.com/snakers4/silero-vad. 2024 (vid. pág. 33).

Tjong Kim Sang, Erik F. “Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition”. En: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 2002 (vid. pág. 54).

Vaswani, Ashish et al. “Attention is All you Need”. En: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017 (vid. págs. 20-22).

Verspoor, Karin y Kevin Bretonnel Cohen. “Natural Language Processing”. En: *Encyclopedia of Systems Biology*. Ed. por Werner Dubitzky et al. New York, NY: Springer New York, 2013, págs. 1495-1498. DOI: 10.1007/978-1-4419-9863-7_158 (vid. pág. 18).

Wang, Changhan et al. “VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation”. En: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, ago. de 2021, págs. 993-1003 (vid. pág. 36).

Yadav, Hemant y Sunayana Sitaram. “A Survey of Multilingual Models for Automatic Speech Recognition”. En: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, jun. de 2022, págs. 5071-5079 (vid. págs. 15, 16).

Zaratiana, Urchade et al. *GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer*. 2023. arXiv: 2311.08526 [cs.CL] (vid. págs. 53, 54).

Zeroual, Imad y Abdelhak Lakhouaja. “Data science in light of natural language processing: An overview”. En: *Procedia Computer Science* 127 (2018). Proceedings of the first international conference on Intelligent Computing in Data Sciences, ICDS2017, págs. 82-91. DOI: https://doi.org/10.1016/j.procs.2018.01.101 (vid. pág. 15).

Zheng, Xianrui, Chao Zhang y Philip C. Woodland. *DNCASR: End-to-End Training for Speaker-Attributed ASR*. 2025. arXiv: 2506.01916 [eess.AS] (vid. pág. 27).