ISBN: 978-628-96613-2-3.

Global Problems, Situated Solutions: The Genesis of an Innovation Project to Counteract Racial and Gender Bias in AI

Milagros Tevez Sauco, Eng.[1], Andrea G. Seminario Vilca, Eng.[2], Celeste H. Bolotra[3], Guadalupe Pascal, MSc.[4]
[1,4] *Universidad del Gran Rosario, Argentina;* [2,3,4] *Universidad Nacional de Lomas de Zamora, Argentina;*
[1,4] *Cátedra Abierta Latinoamericana Matilda y las mujeres en Ingeniería;*
milagros.tevezsauco@gmail.com , andrea.seminario98@gmail.com , celestebolotra@gmail.com , gpascal@ingenieria.unlz.edu.ar

Abstract – This paper presents an open-innovation experience with a gender perspective, aiming to identify and address biases in automatic speech recognition (ASR) systems. The project was developed within the framework of the first Latin American Open Innovation Day, organized by CAL-Matilda, involving faculty and student teams from Argentina. Based on a methodology that combines the innovation cycle and design science, the team identified gender and racial bias in ASR systems, analyzed the composition of existing datasets, evaluated model performance differences, and developed the collaborative platform VocesLatinas.ai. This initiative aims to establish an open repository of Latin American women's voices, thereby enhancing the representativeness of the models used for training. The article documents the analysis, design, and pilot testing processes and reflects on the learnings from a territorial, collaborative, and situated approach.

Keywords– speech recognition, bias, citizen participation, open innovation

Problemas globales, soluciones situadas. Génesis de un proyecto de innovación para revertir sesgos de raza y género en IA

Milagros Tévez Sauco, Ing.[1], Andrea G. Seminario Vilca, Ing.[2], Celeste H. Bolotra[3], Guadalupe Pascal, MSc.[4]
[1,4] *Universidad del Gran Rosario, Argentina;* [2,3,4] *Universidad Nacional de Lomas de Zamora, Argentina;*
[1,4] *Cátedra Abierta Latinoamericana Matilda y las mujeres en Ingeniería;*
milagros.tevezsauco@gmail.com , andrea.seminario98@gmail.com , celestebolotra@gmail.com , gpascal@ingenieria.unlz.edu.ar

Resumen– ***Este trabajo presenta una experiencia de innovación abierta con perspectiva de género orientada a identificar y abordar sesgos en las tecnologías de reconocimiento automático de voz. El proyecto surgió en el marco de la primera Jornada de Innovación Abierta Latinoamericana Matilda (CAL-Matilda), con la participación de equipos docentes y estudiantiles de Argentina. A partir de una metodología basada en el ciclo de innovación y la ciencia de diseño, se identificó el problema del sesgo de género y raza en los sistemas ASR, se analizó la composición de los datasets disponibles, se evaluó el desempeño diferencial de los modelos y se desarrolló la plataforma colaborativa VocesLatinas.ai. Esta iniciativa busca construir un repositorio abierto de voces femeninas latinoamericanas para mejorar la representatividad en los entrenamientos. El artículo documenta el proceso de análisis, diseño y prueba piloto y reflexiona sobre los aprendizajes en clave territorial, colaborativa y situada.***

Palabras clave- reconocimiento de voz, sesgos, participación ciudadana, innovación abierta.

I. Introducción

Las tecnologías de reconocimiento automático de voz (ASR) se han expandido globalmente en aplicaciones cotidianas como asistentes virtuales, dispositivos inteligentes y sistemas de atención automatizada. A pesar de su creciente uso, diversos estudios han evidenciado que estos sistemas no operan con la misma eficacia para todos los grupos sociales. Particularmente, presentan menores niveles de precisión al procesar voces femeninas debido a que han sido entrenados principalmente con registros de voz masculinos. Esta asimetría en la calidad de reconocimiento reproduce, a su vez, un sesgo de desempeño que afecta principalmente a los grupos subrepresentados y refuerza desigualdades preexistentes. [1], [2], [3]

Este sesgo de desempeño, además, implica una paradoja: mientras las voces feminizadas son ampliamente utilizadas como interfaz vocal por su asociación con roles de cuidado y su percepción como "amables" o "cercanas", esos mismos sistemas no logran reconocerlas con igual precisión cuando se trata de emitir un comando o recibir instrucciones. [4], [5], [6]

La problemática se agrava cuando se considera que a estos fenómenos descriptos se les suman otras dimensiones como la raza, el acento o la región, reproduciendo patrones de exclusión que afectan especialmente a poblaciones en vías de desarrollo y con brechas tecnológicas. [7], [8], [9]

A. Contexto del proyecto de innovación y objetivos

En este escenario, este trabajo propone contribuir al debate sobre esta problemática, abordar empíricamente la existencia de estos sesgos y presentar una estrategia de solución a partir de una experiencia de innovación abierta.

El presente trabajo se enmarca en la primera jornada de "Experiencias de Innovación Abierta Latinoamericana Matilda", una iniciativa que busca promover y formar en el ámbito de la innovación abierta con perspectiva de género, orientada a empoderar a jóvenes y mujeres en ingeniería para que desarrollen la capacidad de abordar y resolver problemas significativos en sus territorios. La jornada tuvo por objetivo específico desarrollar habilidades clave de la ingeniería, como el pensamiento creativo, el diseño innovador, la comunicación efectiva y el liderazgo en equipo.

II. Metodología

El enfoque metodológico adoptado para el proyecto combinó principios del ciclo de innovación abierta con la metodología de la ciencia de diseño.

El ciclo de innovación aportó una lógica secuencial basada en la identificación de desafíos, la ideación colaborativa de soluciones y la implementación de prototipos orientados al cambio. [10], [11]

La ciencia de diseño, por su parte, provee herramientas para sistematizar el proceso creativo, ya que enfatiza la creación iterativa de soluciones tecnológicas como respuesta en escenarios reales, priorizando la retroalimentación del entorno y los actores involucrados. [12], [13]

La Figura 1 muestra el esquema metodológico resultante. Esta estructura permite vincular teoría y práctica al mismo tiempo que se incentiva la creatividad y se procuran los ciclos de rigurosidad y relevancia de los desarrollos.

Fig. 1 Esquema metodológico: aportes desde el ciclo de la innovación y ciencias de diseño

La conformación del equipo de trabajo se basó en la convocatoria abierta y voluntaria a la comunidad educativa. Han participado docentes y estudiantes de las cátedras de Planificación, Control y Optimización de Procesos e Investigación Operativa de la Universidad Nacional de Lomas de Zamora, situada en Pcia. de Buenos Aires, Argentina, y la cátedra de Introducción al Aprendizaje Automático de la Universidad del Gran Rosario, situada en Pcia. de Santa Fe, Argentina. El trabajo fue íntegramente en modalidad virtual. Todas las personas del equipo aceptaron ser parte activa de las etapas de trabajo, aplicando técnicas de mentoría entre pares y de trabajo colaborativo.

La selección del desafío a abordar fue resultado de una curaduría sobre problemáticas [14] y desarrollos reales y actuales en el marco de la jornada de experiencias de innovación mencionada previamente. Se desarrolló una preselección de problemáticas en la cual todas ellas estuvieron vinculadas a sesgos de género en modelos algorítmicos y dilemas éticos de la revolución tecnológica actual. El equipo decidió abordar unánimemente la problemática de las tecnologías de reconocimiento de voz por su potencial para articular conocimientos de datos abiertos y territorialidad.

III. Diagnóstico de los sesgos en datos y modelos ASR

Para el análisis empírico del problema, se definieron dos preguntas rectoras, orientadas por la bibliografía de referencia en sesgos algoritmos de reconocimiento de voz:

- ¿Cuál es la diferencia en la precisión de los modelos de reconocimiento automático de voz (ASR) al procesar voces femeninas versus masculinas?
- ¿Qué características de los conjuntos de datos de entrenamiento (como la distribución de género, frecuencia vocal o acento) podrían explicar el sesgo de desempeño observado en modelos de ASR?

Con el objetivo de responder estas preguntas, se llevó a cabo un análisis exploratorio del repositorio Common Voice, una de las plataformas más grandes de datos de voz de acceso público. Common Voice es una colección multilingüe masiva de voces transcritas, desarrollada con un enfoque de ciencia abierta y basada en crowdsourcing tanto para la recolección como para la validación de datos. En esta plataforma, se pueden grabar frases de muestra y validar las grabaciones de otros usuarios. Según la información disponible en el repositorio y estudios previos, la versión de Common Voice más reciente incluye 29 idiomas y ha contado con la participación de más de 50.000 personas, lo que ha generado 2.500 horas de audio. Common Voice es el corpus de audio más grande de dominio público para el reconocimiento de voz, tanto en términos de horas como de idiomas. [15]

A. Análisis de los sesgos en los datos disponibles

La Tabla I presenta la composición del set de datos disponible en Common Voice desde una perspectiva de género e idioma. En el caso de los clips en inglés, el 47% corresponde a voces masculinas, el 19,13% a voces femeninas y el 0,29% a voces no binarias u otras identidades no especificadas. En el caso de los clips en español, las voces masculinas ocupan el 42,50% del corpus, las voces femeninas el 13,9% y las voces no binarias u otras ocupan el 1,77%. [15]

TABLA I
Conformación del conjunto de datos disponible en Common Voice según género (feminino/masculino/other) e idioma (español/inglés)*

Idioma	Clips	Usuarios	Género Masculino	Género Femenino	Otros**
Español	34.76 %	22.21 %	47.57 %	19.13 %	0.29 %
Inglés	65.24 %	77.79 %	42.50 %	13.92 %	1.77 %

*Valores expresados en porcentaje del conjunto de datos disponible.
**Otros géneros no especificados en los metadatos del registro

La Tabla II muestra el análisis por país de cada una de las grabaciones, donde se analizaron las horas de grabación validadas diferenciando entre voces de género masculino y de género femenino, donde es apreciable que el sesgo abarca el 86% de las nacionalidades identificadas en el set de datos.

TABLA II
Conformación del conjunto de datos disponible en Common Voice según género (feminino/masculino) y país*

Top 10 Paises	% Horas validadas masculinas	% Horas validadas Femeninas
Marruecos	59.88%	0.00%
Irán	62.68%	7.56%
Alemania	60.00%	5.58%
Portugal / Brasil	57.63%	5.81%
Etiopía	50.95%	0.00%
Laos	49.98%	0.00%
Afganistán	46.41%	0.00%
Dinamarca	52.90%	7.36%
Rusia / Georgia	43.84%	0.00%
Irak	46.59%	5.03%

Ghana	41.30%	0.00%

*Se muestran los 10 principales países que muestran sesgos en la proporción de grabaciones validadas según los metadatos de Common Voice.

B. Análisis de los sesgos en los datos validados

También se examinó el comportamiento voluntario de validación dentro de la plataforma, considerando la representación demográfica según idioma. Para ello se extrajo una muestra aleatoria representativa y se calculó cuántos registros de audio subidos habían sido efectivamente validados por otros participantes.

Los resultados muestran que, de los 617 registros en inglés analizados, 138 fueron validados (22,3%), mientras que de los 140 registros en español, solo 2 fueron validados (1,4%). Esto sugiere que no solo hay un sesgo en la generación del dato original, sino también en su posterior validación. Esta propagación del sesgo podría afectar significativamente la calidad y diversidad de los sets de datos finales de la plataforma.

C. Evaluación de modelos con datos sesgados.

En vías de cuantificar el sesgo, también se buscó conocer cuál es la diferencia en la precisión de los modelos de reconocimiento automático de voz al procesar distintos géneros e idiomas. Para ello se construyó un conjunto de datos original ad hoc, se seleccionaron muestras del set de Common Voice y se construyó un algoritmo de evaluación [http://voceslatinas.colab] .

Set de datos ad hoc: Se recortaron registros de audio de canciones argentinas y mexicanas, para luego ser transcritos manualmente y de esta forma tener input para entender si el algoritmo transcribe correctamente.

Algoritmo de evaluación: El algoritmo se compone de tres partes principales:

1) Importación de los datos desde la nube. En este paso, el script recibe las grabaciones de audio y las transcripciones validadas, tanto del set ad hoc como de las muestras de Common Voice.
2) Transcripción de los datos a estudiar. El modelo de ASR (Automatic Speech Recognition) utilizado en este paso es Whisper de OpenAI. Este modelo fue elegido por sobre Google Speech-to-Text y Amazon Transcribe por ser gratuito y no requerir entrenamiento o fine-tuning.
3) Evaluación de las grabaciones de audio. Se utilizó como métrica de evaluación el Word Error Rate (WER) importado desde *jiwer* (librería de Python) [16]. Esta métrica se calcula según la siguiente relación:

$$W = \frac{S + D + I}{N}$$

Donde:
S: Sustituciones (palabra incorrecta)
D = Eliminaciones (palabra que falta)
I = Inserciones (palabra agregada)
N = Total de palabras en la referencia

La Tabla III muestra que los registros con voz masculina y en inglés obtuvieron un error promedio de 0,25, con algunos casos en los que el modelo alcanzó una tasa de error igual a cero. En cambio, para los registros con voz femenina o en idioma español, el promedio de error fue superior en más de 10 puntos porcentuales, lo cual refuerza la necesidad de desarrollar datasets más representativos y balanceados para mitigar los sesgos de desempeño.

TABLA III
RESULTADOS DE LA MÉTRICA WORLD ERROR RATE (WER)*

Set de datos	Idioma - Región	Género	WER (med.)	WER (máx.)	WER (mín.)
ad-hoc	Español - Agentina	Fem.	0.29	0.70	0.12
ad-hoc	Español - México	Fem.	0.36	2.13	0.08
CommonVoice	Español - S/D	Masc.	0.38	0.67	0.08
CommonVoice	Inglés - S/D	Masc.	0.25	1.00	0.00

*Análisis comparativo entre un set de datos disponible y abierto vs. set de datos ad-hoc con registros de audio mayoritariamente femeninos y en idioma en español

D. Mapeo de soluciones.

Una vez comprendido el problema, el equipo realizó un mapeo de soluciones ya implementadas en estudios previos. [3], [17], [18] Las estrategias fueron clasificadas en categorías:

- Diversificación de datasets: Incorporar en los datos de entrenamiento una representación equitativa de voces femeninas, masculinas y no binarias, con distintos acentos, edades y tonos.
- Auditorías de sesgo: Establecer mecanismos periódicos para evaluar el desempeño de los modelos con distintos grupos demográficos.
- Técnicas de aprendizaje equitativo (fairness-aware learning): Implementar métodos como la ponderación de datos, penalización por sesgo o estrategias que neutralicen el género como variable predictiva, para mejorar la equidad del modelo.
- Equipos de desarrollo diversos: Promover la conformación de equipos interdisciplinarios con diversidad de género, idioma, cultura e improcedencia territorial, para identificar y mitigar sesgos desde el diseño del artefacto.

Este mapeo permitió establecer una base teórica y práctica sobre la cual se construyó la solución desarrollada por el equipo.

IV. Desarrollo de la innovación: Voceslatinas.ai

Para dar respuesta a la problemática, se creó la plataforma Voceslatinas.ai [http://voceslatinas.ai], una aplicación de acceso abierto con el objetivo de promover la recopilación colaborativa de registros de voces femeninas latinoamericanas, priorizando la diversidad regional, etaria, tonal y de acento. De esta manera, la plataforma busca contribuir a la construcción de datasets más equitativos que puedan ser utilizados en el entrenamiento de modelos de reconocimiento de voz automático, mejorando así su desempeño con poblaciones subrepresentadas. Además, la iniciativa tiene como finalidad sensibilizar sobre los sesgos presentes en la inteligencia artificial y fomentar la participación ciudadana para el desarrollo tecnológico.

A. Características generales del sistema

Voceslatinas.ai es una plataforma que permite recabar, incorporar, almacenar, distribuir y difundir un repositorio de audios y transcripciones de voces femeninas latinas, convirtiéndolos en sets de datos, a partir de la colaboración y curaduría voluntaria de personas sensibilizadas en la temática.

La Figura 2 muestra la arquitectura de funcionamiento de la plataforma. Se puede observar cómo los usuarios interactúan con la interfaz subiendo y descargando audios y CVS con las transcripciones validadas. La información suministrada pasa por la función provincial IA donde transcribe el audio con Whisper, calcula el WER comparando el cvs de la transcripción validada con lo entregado por la IA. Finalmente sube a Supabase la información suministrada junto con los metadatos.

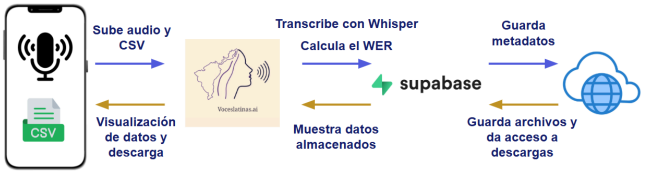

Fig. 2 Arquitectura de Voceslatinas.ai

Esta arquitectura garantiza accesibilidad, trazabilidad, transparencia y promueve la diversidad en la recolección de voces, permitiendo la retroalimentación continua y mejora colaborativa.

B. Lanzamiento y prueba piloto

La aplicación basa su interfaz de usuario en un formulario cuyos campos son: la grabación de voz del usuario, la transcripción correcta de la misma y sus metadatos (género, edad, país y nombre del usuario).

Luego, la aplicación toma los datos cargados, los almacena en Supebase y ejecuta una la evaluación de la grabación utilizando la métrica WER, anteriormente usada para entender el acierto de la transcripción del modelo Whisper.

Finalmente el usuario puede ver el WER entre su audio y transcripción, pudiendo subir nuevas grabaciones con tasas WER menores. De esta forma, el usuario puede contribuir a la generación de datos de calidad que puedan entrenar modelos de forma más eficiente.

V. Conclusiones y discusiones

A lo largo del análisis contextual de la problemática se ha puesto en evidencia que las tecnologías de reconocimiento de voz no están exentas de reproducir sesgos y dilemas éticos característicos del campo de la inteligencia artificial. Estas tecnologías, lejos de ser neutras, reflejan las desigualdades estructurales de los contextos en los que se desarrollan y aplican.

En este sentido, el desarrollo de la plataforma VocesLatinas.ai, basada en una lógica de contribución abierta y colaborativa, apunta a intervenir en dos dimensiones clave. Por un lado, capitaliza el trabajo en red de colegas, estudiantes, docentes y activistas de la región latinoamericana para construir un set de datos diverso, representativo y contextualizado que contribuya a mitigar los sesgos algorítmicos ya identificados. Por otro lado, esta participación voluntaria se propone como una estrategia de apropiación tecnológica y de sensibilización social, promoviendo la comprensión de la relación profunda entre los sesgos de género, raza y territorio y las formas de violencia simbólica que se reproducen a través de las herramientas digitales.

Asimismo, se destaca que la participación en el desafío de innovación abierta impulsado por la CAL-Matilda fue un factor determinante para el surgimiento y consolidación de esta iniciativa. La dinámica propuesta favoreció la cohesión del equipo de trabajo y permitió articular una temática de alto impacto con una metodología rigurosa, alcanzando un resultado concreto, socialmente relevante y tecnológicamente viable.

Este proyecto permite reafirmar la necesidad de desarrollar soluciones con una perspectiva de justicia social. La participación activa, colaborativa y territorializada no solo fortalece la construcción de datos inclusivos, sino que también habilita procesos de empoderamiento tecnológico desde una mirada situada.

Se proyecta continuar con la expansión de la plataforma, potenciar su interoperabilidad y articularla con otras iniciativas regionales orientadas a democratizar el acceso y la participación en el desarrollo de la inteligencia artificial.

Agradecimiento/Reconocimiento

Las autoras agradecen a la CAL-Matilda por promover espacios de innovación abierta con enfoque regional, creativo

y humano en un contexto en el cual los estándares de desarrollo están cada vez más efectivos en términos globales, pero menos eficientes en términos locales.

Referencias

[1] A. Koenecke *et al.*, «Racial disparities in automated speech recognition», *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, n.º 14, pp. 7684-7689, abr. 2020, doi: 10.1073/pnas.1915768117.

[2] R. Tatman, «Gender and Dialect Bias in YouTube's Automatic Captions», en *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, y H. Wallach, Eds., Valencia, Spain: Association for Computational Linguistics, abr. 2017, pp. 53-59. doi: 10.18653/v1/W17-1606.

[3] S. Feng, B. M. Halpern, O. Kudina, y O. Scharenborg, «Towards inclusive automatic speech recognition», *Comput. Speech Lang.*, vol. 84, p. 101567, mar. 2024, doi: 10.1016/j.csl.2023.101567.

[4] B. Tay, Y. Jung, y T. Park, «When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction», *Comput. Hum. Behav.*, vol. 38, pp. 75-84, 2014, doi: 10.1016/j.chb.2014.05.014.

[5] N. Ni Loideain y R. Adams, «From Alexa to Siri and the GDPR: The Gendering of Virtual Personal Assistants and the Role of EU Data Protection Law», 9 de noviembre de 2018, *Social Science Research Network, Rochester, NY*: 3281807. doi: 10.2139/ssrn.3281807.

[6] N. Ni Loideain y R. Adams, «Female Servitude by Default and Social Harm: AI Virtual Personal Assistants, the FTC, and Unfair Commercial Practices», 11 de junio de 2019, *Social Science Research Network, Rochester, NY*: 3402369. doi: 10.2139/ssrn.3402369.

[7] C. Nass, Y. Moon, y N. Green, «Are machines gender neutral? Gender-stereotypic responses to computers with voices», *J. Appl. Soc. Psychol.*, vol. 27, n.º 10, pp. 864-876, 1997, doi: 10.1111/j.1559-1816.1997.tb00275.x.

[8] A. Kaplan y M. Haenlein, «Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence», *Bus. Horiz.*, vol. 62, n.º 1, pp. 15-25, ene. 2019, doi: 10.1016/j.bushor.2018.08.004.

[9] A. Schlesinger, K. P. O Hara, y A. Taylor, «Lets Talk about Race: Identity, Chatbots, and AI», en *CHI '18*, New York, USA: ACM Press, abr. 2018. doi: 10.1145/3173574.3173889.

[10] G. Mulgan, «Innovation in the public sector: How can public organisations better create, improve and adapt», *Lond. Nesta*, vol. 11, 2014, Accedido: 15 de agosto de 2025. [En línea]. Disponible en: https://scholar.google.com/scholar?cluster=5554698364380966373&hl=en&oi=scholarr

[11] G. Mulgan, «Design in public and social innovation: what works and what could work better», *Nesta Lond.*, 2014, Accedido: 15 de agosto de 2025. [En línea]. Disponible en: https://scholar.google.com/scholar?cluster=14330432933016357503&hl=en&oi=scholarr

[12] A. R. Hevner, S. T. March, J. Park, y S. Ram, «Design Science in Information Systems Research», mar. 2004.

[13] K. Peffers, T. Tuunanen, M. A. Rothenberger, y S. Chatterjee, «A Design Science Research Methodology for Information Systems Research», *J. Manag. Inf. Syst.*, vol. 24, n.º 3, pp. 45-77, dic. 2007, doi: 10.2753/MIS0742-1222240302.

[14] C. Criado-Perez y A. Echevarría, *La mujer invisible: descubre cómo los datos configuran un mundo hecho por y para los hombres*. Seix Barral, 2020. Accedido: 15 de agosto de 2025. [En línea]. Disponible en: https://dialnet.unirioja.es/servlet/libro?codigo=835483

[15] R. Ardila *et al.*, «Common Voice: A Massively-Multilingual Speech Corpus», 5 de marzo de 2020, *arXiv*: arXiv:1912.06670. doi: 10.48550/arXiv.1912.06670.

[16] A. C. Morris, V. Maier, y P. Green, «From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition», presentado en Proc. Interspeech 2004, 2004, pp. 2765-2768. doi: 10.21437/Interspeech.2004-668.

[17] R. Tatman, «Why ASR + NLP isn't enough for commercial language technology», *J. Acoust. Soc. Am.*, vol. 150, n.º 4_Supplement, p. A347, oct. 2021, doi: 10.1121/10.0008537.

[18] Y. Peng *et al.*, «NTU Speechlab LLM-Based Multilingual ASR System for Interspeech MLC-SLM Challenge 2025», 4 de julio de 2025, *arXiv*: arXiv:2506.13339. doi: 10.48550/arXiv.2506.13339.