Este artículo se encuentra disponible
en acceso abierto bajo la licencia Creative
Commons Attribution 4.0 International License

REVISTA OFICIAL DEL PODER JUDICIAL

Vol. 16, n.º 21, enero-junio, 2024, 53-81
ISSN: 2663-9130 (En línea)
DOI: 10.35292/ropj.v16i21.881

Comparación forense de voces: un estudio preliminar sobre las diferencias entre una voz natural y una voz artificial para la investigación judicial[1]

Forensic comparison of voices: a preliminary study on the differences between a natural voice and an artificial voice for judicial investigation

Comparação forense de vozes: um estudo preliminar sobre as diferenças entre voz natural e voz artificial para investigação judicial

JHON JIMENEZ PEÑA
Universidad Nacional Mayor de San Marcos
(Lima, Perú)
Contacto: jhon.jimenez@unmsm.edu.pe
https://orcid.org/0000-0003-3317-6152

FERNANDO AARÓN TORRES CASTILLO
Universidad Nacional Mayor de San Marcos
(Lima, Perú)
Contacto: fernando.torres2@unmsm.edu.pe
https://orcid.org/0000-0002-1432-8811

OSCAR ESAUL CUEVA SANCHEZ
Universidad Nacional Mayor de San Marcos
(Lima, Perú)
Contacto: oscar.cueva1@unmsm.edu.pe
https://orcid.org/0000-0003-1361-2367

1 Esta investigación fue impulsada por el Gabinete de Lingüística Forense —del Instituto de Investigación de Lingüística Aplicada (CILA) de la Universidad Nacional Mayor de San Marcos—, el cual fue creado mediante la Resolución Decanal n.º 000623-2021-D-FLCH/UNMSM.

RESUMEN

Este estudio presenta una aproximación en torno a las similitudes y las diferencias fonéticas entre una voz natural y una voz artificial, por lo que se busca: (a) brindar un análisis que sirva de antecedente ante casos judiciales de clonación de voz por inteligencia artificial (IA) y (b) exponer la importancia de la lingüística como fuente de evidencia científica para el sistema judicial. Así, se ha analizado la voz del narrador argentino Mariano Closs y su contraparte artificial creada en FakeYou (convertidor de texto en habla) mediante el método combinado que integra el uso de programas automáticos de análisis de voz (Forensia y SIS II) y el análisis fonético. Los programas automáticos mostraron resultados de alta convergencia entre la voz natural y la voz artificial. Sin embargo, en el análisis fonético, se observó diferencias en la producción de determinados sonidos, en la entonación; asimismo, hubo procesos fonéticos presentes en una muestra. Es así que, a pesar de la similitud de las muestras en el plano biométrico, la voz artificial del narrador Mariano Closs aún no es del todo similar a su contraparte natural en el plano fonético.

Palabras clave: análisis fonético; voz artificial; convertidor de texto en habla; lingüística forense; criminalística.

Términos de indización: fonética; habla; lingüística; procedimiento legal; crimen (Fuente: Tesauro Unesco).

ABSTRACT

This study presents an approach to the phonetic similarities and differences between a natural voice and an artificial voice, which is why it seeks to: (a) provide an analysis that serves as a precedent for judicial cases of voice cloning by artificial intelligence (AI) and (b) expose the importance of linguistics as a source of scientific evidence for the judicial system. Thus, the voice of the Argentine narrator Mariano Closs and his artificial counterpart created in FakeYou (text-to-speech converter) have been analyzed using the combined method that integrates the use of automatic voice analysis programs (Forensia and SIS II) and the phonetic analysis. The automatic programs showed results of high convergence between the natural voice and the artificial voice. However, in the

phonetic analysis, differences were observed in the production of certain sounds, in intonation, and there were phonetic processes present in a sample. Thus, despite the similarity of the samples on the biometric level, the artificial voice of the narrator Mariano Closs is still not completely similar to his natural counterpart on the phonetic level.

Key words: phonetic analysis; artificial voice; text to speech converter; forensic linguistics; criminalistics.

Indexing terms: phonetics; speech; linguistics; judicial procedure; crime (Source: Unesco Thesaurus).

RESUMO

Este estudo apresenta uma abordagem sobre as semelhanças e diferenças fonéticas entre uma voz natural e uma voz artificial, por isso busca: (a) fornecer uma análise que sirva de precedente para casos judiciais de clonagem de voz por inteligência artificial (IA) e (b) expor a importância da linguística como fonte de evidências científicas para o sistema judicial. Assim, a voz do narrador argentino Mariano Closs e sua contraparte artificial criada no FakeYou (conversor de texto para fala) foi analisada através do método combinado que integra o uso de programas de análise automática de voz (Forensia e SIS II) e a análise fonético. Os programas automáticos apresentaram resultados de alta convergência entre a voz natural e a voz artificial. Porém, na análise fonética foram observadas diferenças na produção de determinados sons, na entonação, e houve processos fonéticos presentes em uma amostra. Assim, apesar da semelhança das amostras no nível biométrico, a voz artificial do narrador Mariano Closs ainda não é totalmente semelhante à sua contraparte natural no nível fonético.

Palavras-chave: análise fonética; voz artificial; conversor de texto para fala; linguística forense; criminalística.

Termos de indexação: fonética; fala; linguística; procedimento legal; crime (Fonte: Unesco Thesaurus).

Recibido: 18/10/2023 **Revisado:** 25/10/2023
Aceptado: 7/5/2024 **Publicado en línea:** 30/6/2024

1. INTRODUCCIÓN

Según Ramírez (2023) —redactor del periódico *El Comercio*—, hasta septiembre del presente año se han identificado, por lo menos, cincuenta casos de clonación de voz con inteligencia artificial (IA) para estafar o fingir secuestros. Este tipo de casos poco a poco va tomando terreno en el Perú, por ello es importante realizar estudios que examinen las diferencias entre la voz natural y la voz artificial para así brindar un antecedente a los especialistas que analicen estos casos.

La inteligencia artificial pertenece a una rama de la ciencia computacional que tiene como meta generar procesos cognitivos similares a los de los humanos (Peña, 2022). Además, cumple un rol importante en la actualidad porque se usa en diversos ámbitos y para distintos fines (organización de bases de datos, procesos logísticos, asistentes virtuales, replicación y creación de voz, entre otros). De esta forma, la inteligencia artificial en su propósito de igualar a la competencia del lenguaje humano ha tenido avances muy significativos.

Es así que actualmente existen muchos conversores de texto en habla a disposición de cualquier persona, por lo que ahora es muy común ver en redes sociales diversos contenidos en los que se usan voces artificiales e, incluso, puede resultar complicado distinguir cuándo se trata de una voz natural o una voz artificial. En ese sentido, la lingüística cumple un rol importante porque «todo texto oral o escrito involucrado en delitos tipificados en el Código Penal —es decir, que es utilizado en la investigación fiscal y empleado en la administración de justicia [...]— es potencialmente objeto de estudio de la lingüística forense» (Lazo y Rivas, 2022, p. 374). Y aunque desde la lingüística aún no se ha estudiado a profundidad casos como el de la clonación de voz

> [se considera] que tiene un gran potencial a la hora de abordar uno de los desafíos de seguridad más importantes que enfrenta el mundo en la actualidad. Nos referimos a los *deepfakes*, [son] vídeos o audios que, sin ser reales, lo parecen debido a una manipulación [...], realizada mediante técnicas de inteligencia artificial. (San Segundo, 2022)

El desarrollo de estas nuevas tecnologías puede tener distintos efectos tanto positivos, por ejemplo, *Illariy*, quien es la primera presentadora de noticias generada por inteligencia artificial que habla en quechua (un proyecto desarrollado desde la Facultad de Letras y Ciencias Humanas de la Universidad Nacional Mayor de San Marcos), pero los efectos también pueden ser negativos si estas tecnologías se usan de forma equivocada. Por ejemplo, no sería raro que los delitos de fraude, usurpación de la identidad, extorsión, amenazas, violación de la privacidad por clonación de voz aumenten porque cada vez es más accesible replicar voces. Uno de los convertidores de texto en habla más famosos, justamente por ser de uso libre, es el sitio web FakeYou, el cual permite al usuario convertir un texto en habla con la voz de una celebridad o cualquier personaje que esté en su base de datos e, incluso, replicar la voz de cualquier persona siempre que se realice una suscripción y se cuente con grabaciones de audio de la voz que se busca replicar. No obstante, a pesar de que el sitio web advierta lo siguiente: «No aprobamos el uso de FakeYou para ningún tipo de suplantación, engaño, insulto, abuso o maltrato de cualquier grupo» (Echelon, s. f.), es inevitable que dicho convertidor de texto en habla pueda ser usado con fines delictivos.

Por esta razón, el objetivo del presente estudio es brindar una aproximación en torno a las similitudes y las diferencias fonéticas entre una voz natural y una voz artificial para así poder identificar parámetros en los que ambas voces difieran. Para realizar esta tarea, se analizó la voz de Mariano Closs, relator argentino y periodista deportivo (la muestra se extrajo de entrevistas encontradas en internet) y su contraparte artificial (proveniente de FakeYou). Es importante señalar que el estudio se circunscribe al campo de la fonética forense, puesto que se realiza la comparación de voces entre ambas muestras. Asimismo, el análisis se realiza con el método combinado que integra el uso de programas automáticos de análisis de voz (Forensia y SIS II) y el análisis fonético del habla.

El presente artículo se estructura en cinco apartados. En el primer apartado, se contextualiza e identifica el problema del estudio. En el segundo apartado, se expone el marco teórico. En el tercer apartado, se presenta la metodología de recolección y acondicionamiento de los

datos. En el cuarto apartado, se presenta el análisis automático y fonético. Finalmente, en el último apartado, se presentan las conclusiones de la investigación.

2. MARCO CONCEPTUAL

La lingüística forense se nutre de campos como la fonética, la fonología, la sociolingüística, entre otras ramas, con la finalidad de esclarecer un hecho delictivo. El presente estudio se circunscribe en el campo de la fonética forense y la conversión de texto en habla.

2.1. Fonética forense

La fonética se encarga de la descripción de los sonidos del habla desde tres perspectivas: articulatoria, acústica y sonora (Garayzábal *et al.*, 2019). Mientras que la fonética forense se define como «principalmente el uso de técnicas fonéticas en el análisis de la voz aplicado a investigaciones criminales. Incluye técnicas de comparación de voz, reconocimiento de voz […]» (Olsson, 2008, p. 156).

En ese sentido, la relación de la fonética forense y la criminalística es muy estrecha porque el interés fundamental de la fonética forense «reside en discernir con el mayor grado de fiabilidad posible si concurren suficientes indicios como para sostener que dos voces pueden corresponder a la misma persona o si, por el contrario, hay que rechazar esta posibilidad» (Fernández, 2007, p. 49). Es importante señalar que la voz puede variar a nivel idiolectal, lo que se conoce como «el uso individual [de la lengua] que establece un hablante y que diversos factores como los culturales, económicos, educativos, sociales, de género u profesión se manifiestan en estos idiolectos» (Torres, 2023, p. 15) y a nivel interhablante (variación interhablante).

En la comparación de locutores, se usan distintos tipos de análisis. Gold y French (2011) enlistan los métodos usados en distintos países: análisis fonético auditivo (AuPA), análisis fonético acústico (AcPA), análisis fonético auditivo y fonético acústico (AuPA + AcPA), análisis por sistema automático de reconocimiento de voces (ASR) —uso de *softwares* biométricos de análisis de voz— y, finalmente, el análisis por

sistema automático de reconocimiento de voces con el análisis humano (HASR) —combina todos los métodos anteriores—.

2.2. La conversión de texto en habla

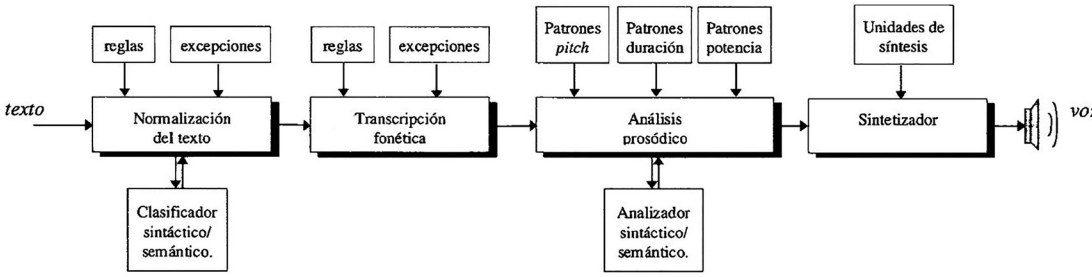
Cortez *et al.* (2009) señalan que «una de las tareas fundamentales de la inteligencia artificial (IA) es la manipulación de lenguajes naturales usando herramientas de computación, [… los lenguajes de programación] forman el enlace necesario entre los lenguajes naturales y su manipulación por una máquina» (pp. 47-48). Y, justamente, los convertidores de texto en habla son el resultado del procesamiento del lenguaje natural a través de un lenguaje de programación que se estructura modularmente.

En ese sentido, Bonafonte (1997) describe la conversión de texto en voz como un «sistema que requiere unos registros de señales orales relacionadas con unas unidades básicas (por ejemplo fonemas), que ha de concatenar siguiendo el texto de entrada» (p. 70). Además, señala que, para obtener una conversión de calidad, las unidades básicas deben ser modificadas para que se reproduzcan de la forma más natural posible. Añade que el análisis de los sonidos y su interacción, tanto como los patrones suprasegmentales son tarea de fonetistas y lingüistas.

En la figura 1, se detalla el funcionamiento esencial de la conversión de texto en voz.

Figura 1

Sistema de conversión de texto en voz propuesto por Bonafonte



Nota. Tomado de Bonafonte (1997, p. 71).

Según Llisterri *et al*. (2004), la conversión de texto en habla permite que cualquier texto escrito sea oralizado por un ordenador con la ayuda de una serie de módulos que procesan los datos de naturaleza lingüística y, asimismo, recurren también a bases de datos que contienen información de ese tipo. Y aunque este campo tradicionalmente se ha asociado con la ingeniería de telecomunicaciones y el tratamiento digital de señales, en la actualidad es necesaria también la participación de expertos que faciliten el conocimiento lingüístico en cada uno de los módulos.

Fernández (2007) indica que hay tres tipos principales de síntesis de voz: la síntesis por formantes, la síntesis articulatoria y la síntesis por concatenación. La síntesis por formantes genera el habla a partir de la especificación previa de los parámetros acústicos, la síntesis articulatoria genera el habla a partir de parámetros que describen la posición y el movimiento de los articuladores y la síntesis por concatenación genera el habla uniendo pequeños fragmentos de sonido para generar oraciones. En el caso de FakeYou, este es un convertidor de texto en habla que permite acceder a una gran cantidad de voces de su base de datos (voces de deportistas, presentadores, actores, etc.). Además, con una suscripción permite generar la voz artificial de cualquier persona a partir de archivos de audio que previamente se carguen en el sitio web.

2.3. Funcionamiento de las redes neuronales artificiales

Según Mena y Rojas (2021), para que una inteligencia artificial clone la voz humana, la inteligencia artificial debe utilizar modelos que identifiquen patrones, sonidos, estructuras silábicas, palabras, entre otros elementos de la voz humana, por ende, utilizan redes neuronales computacionales, estas

> buscan simular la forma en la que el cerebro humano es capaz de reconocer la voz y las palabras del entorno que lo rodea, tiene la capacidad de ajustarse a sí misma y mejorar sus resultados conforme transcurre el tiempo [...]. (p. 88)

3. METODOLOGÍA

Este estudio es de tipo exploratorio, puesto que la problemática que se abordará ha sido poco estudiada y no se encontraron antecedentes directos que se refieran al tema. Además, se busca comparar la voz natural con la voz artificial para identificar rasgos fonéticos que ayuden a distinguir una muestra de la otra.

El método que se usa para la presente investigación es el método combinado (Univaso, 2016). Este se compone por el método por sistemas de reconocimiento automático y por el método clásico, el primero se relaciona con el uso de programas de biometría de voz que comparan automáticamente una muestra dubitada con una muestra indubitada a partir de algoritmos (Garayzábal *et al.*, 2019) y, el segundo, se enfoca en un análisis perceptual y acústico mediante la escucha y la visualización del espectrograma y el oscilograma de procesos fonético-fonológicos.

3.1. Recolección de datos

El estudio emplea dos muestras: la muestra natural y la muestra artificial de Mariano Closs. Es importante recalcar que la voz natural de Mariano Closs se extrae de relatos de partidos de fútbol, debido a que la voz artificial proveniente de FakeYou se encuentra también en dicho contexto.

3.1.1. Voz natural

La muestra de voz natural está constituida por dos archivos de audio provenientes de videos de YouTube en los que Mariano Closs narra partidos de la Champions League temporada 2021-2022, ambos videos fueron tomados de la cuenta de YouTube ESPN Fans (en las referencias se especifican los videos). Asimismo, es importante señalar que para descargar los videos en archivos de audio en formato wav se usó el convertidor en línea *y2mp3.top* (https://y2mp3.top). En la figura 2, se reportan las propiedades de ambos archivos de audio:

Figura 2

Archivos que constituyen la muestra de voz natural

3.1.2. Voz artificial

La muestra de voz artificial está constituida también por dos archivos de audio provenientes del sitio web FakeYou (https://fakeyou.com). Este sitio web posee tres opciones de voz para Mariano Closs, se usó: «Mariano Closs (Relator de fútbol Argentino) (por Vox_Populi)». En la figura 3, se muestra una captura de pantalla del sitio web en el que se observa uno de los corpus usados.

Figura 3

Interfaz de FakeYou

El corpus que se introdujo en FakeYou se constituyó por las mismas oraciones y las frases presentes en la muestra natural. La muestra de mayor duración se usó para el análisis fonético. Mientras que la segunda muestra se utilizó para el análisis con los programas automáticos.

Figura 4

Archivos que constituyen la muestra de voz artificial

3.2. Acondicionamiento de las muestras

3.2.1. Voz natural

3.2.1.1. Formato de las muestras de voz natural

Los audios de la muestra de voz natural se descargaron en formato wav. Estos archivos de audio presentan un solo canal, una frecuencia de muestreo de 44.1 kHz y 16 bits de profundidad. Estas características son compatibles con los programas utilizados en el análisis de voz.

3.2.1.2. Extracción de los fragmentos de voz de las muestras de voz natural

En cada archivo de audio, se seleccionaron los intervalos de tiempo correspondientes a la participación de la voz de Mariano Closs. Posteriormente, estos intervalos de tiempo fueron extraídos y concatenados y se crearon nuevos archivos de audio a partir de estos, los mismos que se presentan en la tabla 1, junto a su tiempo de duración. Además, en esta tabla aparece el código de referencia, que será utilizado en adelante para señalar los audios de voz natural.

Tabla 1

Valores temporales de los intervalos de tiempo de la muestra de voz natural

Referencia	Nombre del archivo	Duración (s)
NV_01	¡BENZEMA BRILLÓ Y EL MERENGUE BORRÓ AL PSG DE MESSI DE LA CHAMPIONS!-C.wav	287.56
NV_02	¡ÉPICA REMONTADA E HISTÓRICA CLASIFICACIÓN DEL MERENGUE!-C.wav	210.35

3.2.2. Voz artificial

3.2.2.1. Formato de las muestras de voz artificial

El sitio web FakeYou exporta los archivos de audio en formato wav. Estos archivos de audio presentan un solo canal, una frecuencia de muestreo de 32 kHz y 16 bits de profundidad. Estas características son compatibles con los programas utilizados en el análisis de voz, por lo que no se les realizó ninguna adaptación.

3.2.2.2. Extracción de los fragmentos de voz de las muestras de voz artificial

No fue necesario extraer los fragmentos de habla del audio de la voz artificial, debido a que contenía únicamente la voz artificial de Mariano Closs. En la tabla 2, se muestra la duración de los archivos de audio, así como su código de referencia:

Tabla 2

Valores temporales de la muestra de voz artificial

Referencia	Nombre del archivo	Duración (s)
AV_01	av_mc1.wav	147.77
AV_02	av_mc2.wav	35.05

Cabe señalar que, como se indicó anteriormente, AV_01 se usó en el análisis fonético, mientras que la muestra AV_02 se usó en el análisis automático.

4. ANÁLISIS

4.1. Análisis de comparación automática

En este apartado, se usaron dos programas de comparación automática de voz en sus versiones de prueba o demo con la finalidad de observar los análisis y los resultados realizados por estos.

Los resultados de estos programas están basados en el marco de *likelihood ratio* (razón de verosimilitud o relación de verosimilitud). La introducción del enfoque del LR (o enfoque bayesiano) en las ciencias forenses, incluyendo la comparación de voz, es probabilístico y es denominado como el nuevo paradigma o cambio de paradigma y «muchos estadísticos forenses [lo] recomiendan como el marco lógicamente correcto para la evaluación de las evidencias comparativas» (Morrison, 2009/2011, p. 5).

La razón de verosimilitud se expresa a través del cálculo entre dos probabilidades: las que proceden del mismo locutor y las que proceden de diferentes locutores. En la figura 5, se detalla este cálculo, LR (*likelihood ratio*) es la razón de verosimilitud; p es probabilidad; H_p es la hipótesis del fiscal y H_d es la hipótesis de la defensa. Se debe agregar que el numerador de la fórmula es una expresión de la similitud, y el denominador, una expresión de la tipicidad (Morrison, 2009/2011). Finalmente, si el valor del cálculo de la fórmula es mayor que uno, indica que ambas voces proceden del mismo locutor (hipótesis fiscal); por su lado, si el resultado de la fórmula es menor que uno, indica que las voces proceden de diferentes locutores (hipótesis de la defensa).

Figura 5

Fórmula según el teorema de Bayes

$$LR = \frac{p\,(E \mid H_p)}{p\,(E \mid H_d)}$$

Nota. Tomado de Rosas *et al.* (2011, p. 20).

4.1.1. Comparación automática en SIS II (versión trial)

El *software* SIS II fue creado por la empresa rusa Speech Technology Center (STC) y actualmente se utiliza en diferentes países como México, Perú, Rusia, entre otros. Algunos de los estudios que emplean este programa son los de Jimenez *et al.* (2022) y Torres (2023), ambos autores concluyen que el programa SIS II es óptimo para determinar la convergencia entre las muestras utilizadas.

El programa incorpora diversos *plugins*, el que se utiliza en el presente estudio es el de *Automatic Comparison*, el cual permite comparar dos muestras de voz a partir de tres métodos: SF (método de identificación de espectro de formantes o Spectral-Formant Speaker Identification Method), Pitch (método de identificación estadístico del Pitch o Pitch Statistics Identification) y GMM (modelo mixto gaussiano o método de variabilidad total o Total Variability Method). Con esto se logra determinar si las muestras comparadas pertenecen a un mismo locutor o a diferentes locutores. El método SF se basa en el estudio estadístico de la frecuencia de los formantes para determinar la geometría del tracto vocal, el cual es único para cada persona. El método Pitch es un análisis estadístico de dieciséis características de la frecuencia fundamental (correlato acústico de la vibración de los pliegues vocálicos). El método GMM o variabilidad total (TotV) es un análisis de la representación espectral de una señal de voz a través de coeficientes cepstrum. Posteriormente, la densidad de la distribución de las características de identificación se moldea a través de los modelos de mezclas gaussianas (GMM).

Los resultados del programa SIS II muestran los valores de FR (falso rechazo), FA (falsa aceptación) y LR (*likelihood ratio* o razón de verosimilitud), además, concluye en términos de *same speaker* (mismo locutor) o *different speakers* (locutores diferentes), cada uno con nivel de probabilidad: *high* (alta), *medium* (media) y *low* (baja).

En las figuras 6 y 7, se presentan los resultados de la comparación automática de las muestras del estudio. Como se observa, en ambos casos los valores del LR superan el valor de uno (4715 y 5001 en la figura 6 y 7, respectivamente), lo que indica que se trata del mismo locutor (*same speaker*) con una probabilidad alta (*high probability*).

Figura 6

Comparación automática de la muestra NV_01 y av_mc-C

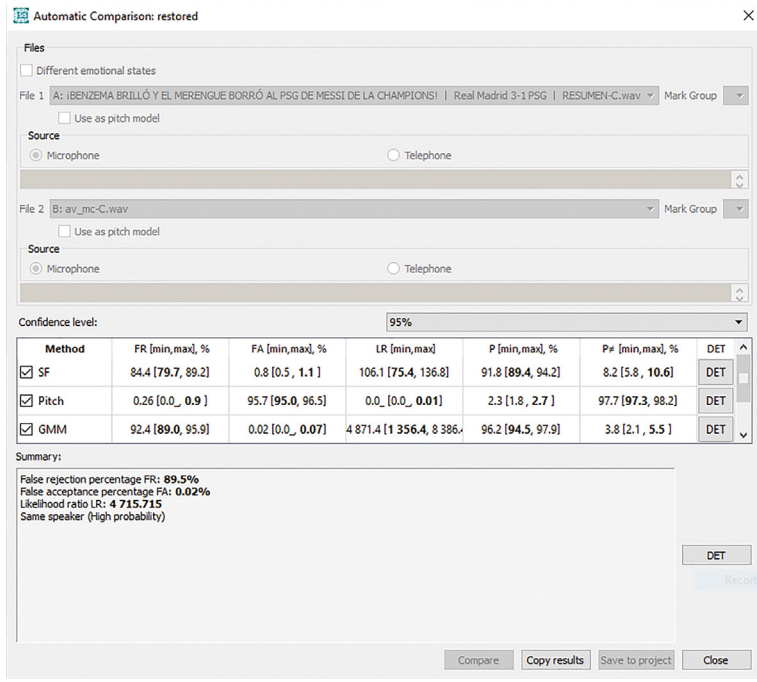
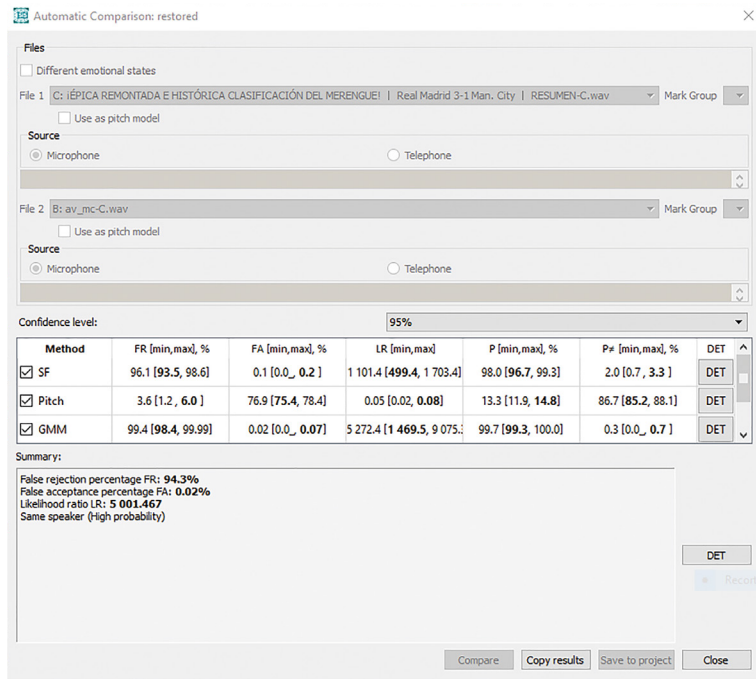


Figura 7

Comparación automática de la muestra NV_02 y av_mc-C



Los resultados indican que la voz natural y la voz artificial de Mariano Closs pertenecen al mismo locutor. No obstante, si se observan los resultados del LR en cada parámetro, se aprecia que el LR en el Pitch tiene un valor menor a uno en ambas figuras (0.0 y 0.05 en la figura 6 y 7, respectivamente), lo que indica que en este parámetro se presentaron diferencias entre locutores —aunque no las suficientes como para determinar el resultado de la comparación automática—, este puede ser un parámetro a tener en cuenta cuando se analiza una voz natural con una voz artificial.

4.1.2. Forensia (versión demo)

Forensia es un programa argentino de comparación automática de voz creado por BlackVOX[2]. Las características discriminatorias de las muestras de voz (dubitada e indubitada) se extraen a partir del análisis espectral [que no considera el Pitch], además, se evalúan con una base de datos universal (Universal Background Model [UBM]) que incluye voces de diferentes lugares de Argentina (Univaso *et al*., 2020).

Asimismo, incluye una etapa de calibración PLDA (*Probabilistic Linear Discriminant Analysis*), que resuelve los inconvenientes relacionados con que las muestras provengan de diferentes canales. Los resultados están presentados en términos de LR en escala logarítmica LLR, en la tabla 3, se presentan las equivalencias entre LR y LLR.

Tabla 3

Equivalencias entre LR y LLR

LR	LLR	Escala verbal	Apoyo
> 10000	> 4	Muy fuerte	A la hipótesis fiscal
1000 a 10000	3 a 4	Fuerte	
100 a 1000	2 a 3	Moderadamente fuerte	
10 a 100	1 a 2	Moderada	
1 a 10	0 a 1	Limitada	

2 Acceso a la página web a través de https://blackvox.com.ar/

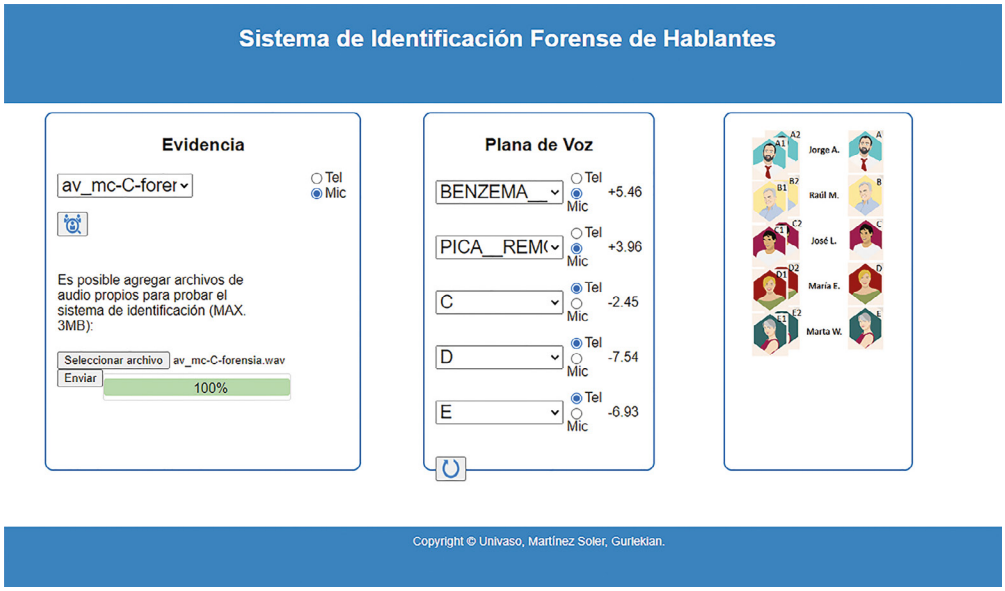
1 a 0.1	0 a -1	Limitada	A la hipótesis de la defensa
0.1 a 0.01	-1 a -2	Moderada	
0.01 a 0.001	-2 a -3	Moderadamente fuerte	
0.001 a 0.0001	-3 a -4	Fuerte	
<0.0001	>-4	Muy fuerte	

Nota. Adaptado de Rosas *et al.* (2011, p. 23).

En la figura 8, se presentan los resultados del programa Forensia en su versión demo[3]. Se observa que en la comparación de la muestra de voz artificial con las dos muestras de voz natural, el valor del LLR es mayor a uno (5.46 y 3.96 en la comparación de NV_01 y NV_02, respectivamente). Esto indica que el peso de la evidencia es muy fuerte y apoya a la hipótesis fiscal.

Figura 8

Comparación entre la muestra artificial y natural



3 El acceso fue permitido por los doctores Pedro Univaso y Jorge Gurlekian durante el curso Identificación por Voz que se imparte en la Universidad Tecnológica Nacional (Buenos Aires), en el año 2022.

Finalmente, la comparación entre la voz artificial y las voces de los locutores C, D y E (voces por defecto en la versión demo) presenta resultados de LLR negativos (menores a 1) cuya fuerza oscila entre moderada y muy fuerte para el apoyo a la hipótesis de la defensa. Asimismo, se debe advertir que al usar la versión demo no se cuenta con todas las características y las funciones del programa.

4.2. Análisis fonético

Machuca *et al.* (2014) y Lazo (2023) señalan que para realizar un análisis de comparación adecuado se debe observar las características fonético-acústicas de las muestras. Es así que «[...] el análisis auditivo, complementado con un análisis acústico detallado, permite establecer los fenómenos que se han de contemplar en el informe pericial sobre las grabaciones dubitadas» (Machuca *et al.*, 2014, p. 100).

Para realizar el análisis fonético, se usó el programa Praat, que ofrece una amplia gama de funciones para explorar y comprender los aspectos acústicos y fonéticos del habla (Boersma y Weenink, 2023). Además, Praat genera espectrogramas y oscilogramas, los cuales son esenciales para el análisis de los sonidos y los patrones fonéticos. Cabe señalar que los dibujos de los espectros se realizaron con el *plugin* TgDraw [Praat plug-in] Versión 0.3 (Muñoz, 2020).

En la tabla 4, se observan algunas de las características fonéticas más resaltantes de la muestra. Como se puede advertir, hay características que indican que la IA que produce la voz artificial de Mariano Closs aún es imperfecta porque se encuentran errores en cuanto a la correspondencia de grafías, la inserción de vocales en ciertos contextos en los que no debería suceder. Asimismo, se reportaron algunos procesos interesantes de observar porque parecen ser netamente lingüísticos, aunque no se descarta que puedan ser perfeccionados con una programación más minuciosa de la IA. Es importante recalcar que los contextos de análisis fueron similares en ambas muestras.

Tabla 4

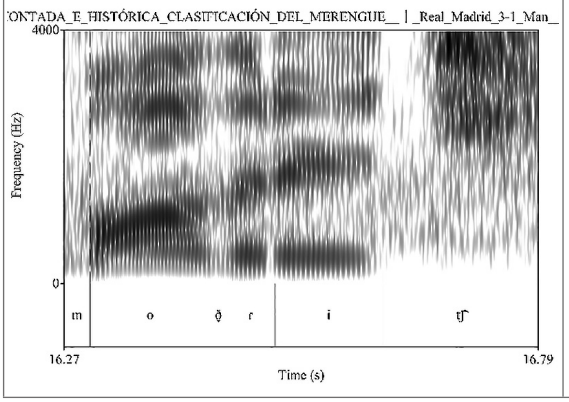
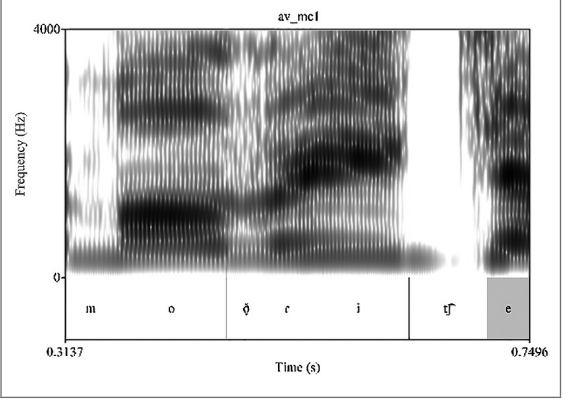
Tabla de comparación de procesos fonéticos de las muestras

	NV	AV
	Producción sin epéntesis	Epéntesis después de la africada palatal a final de palabra
(1)	7.63 – 8.08 Modrich [moðɾitʃ] 16.26 – 16.84 Modrich [moðɾitʃ]	0.31 – 0.74 Modrich [moðɾitʃe] 15.10 – 15.80 Modrich [moðɾitʃe]
	Producción sin elisión	Elisión en la secuencia vocálica -ao a final de palabra
(2)	16.70 – 17.44 Militao [militao] 46.52 – 47.06 Militao [militao]	3.10 – 4.37 Militao [milito:] 9.47 – 10.70 Militao [milito:]
	Producción palatal de la nasal palatal	Producción alveolar de la nasal palatal
(3)	121.25 – 121.61 señoras [seɲoɾas]	5.94 – 6.45 señoras [senoɾas]
	Producción múltiple de la vibrante múltiple	Producción aproximante de la vibrante múltiple
(4)	36.09 - 36.51 Torre [tore] 108.25 - 108.80 remate [remate]	52.20 - 52.73 Torre [toɹe] 31.81-32.22 remate [ɹemate]
	La entonación tiene un rango variable en el alargamiento vocálico	La entonación tiene un rango estático en el alargamiento vocálico
(5)	74.69 – 77.71 señoras y señores, va contra Carlos Vio: 258.66 – 259.92 va a sacar Ederson	0.03 – 3.02 señoras y señores, va contra Carlos Vio: 0.02 – 1.37 va a sacar Ederson

El primer proceso fonético (1) es un caso de epéntesis después de la africada palatal a final de palabra. Como se observa en la tabla 5, en la muestra artificial se produce la inserción de la vocal [e] a final de palabra en *Modrić*, mientras que en la muestra natural la producción se realiza sin ese proceso.

Tabla 5

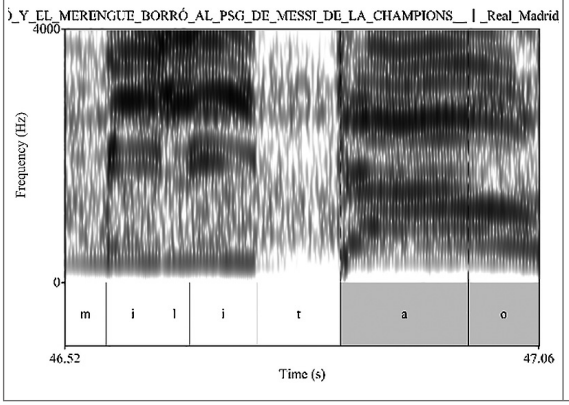
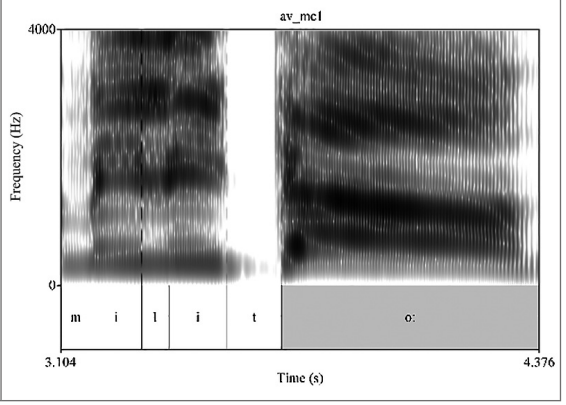
Tabla de rasgo (1)

Natural voice	***Artificial voice***
	

Por su lado, en la tabla 6, se observa un proceso de elisión en la secuencia vocálica -ao a final de palabra. Se aprecia que en la muestra artificial hay una caída de la vocal [a], no obstante, la duración no parece ser afectada porque la vocal dura alrededor de 36 ms.

Tabla 6

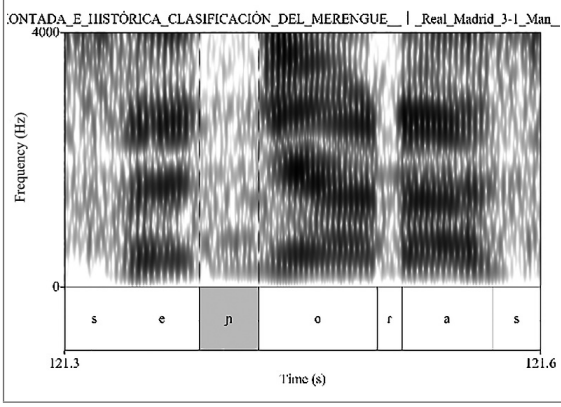
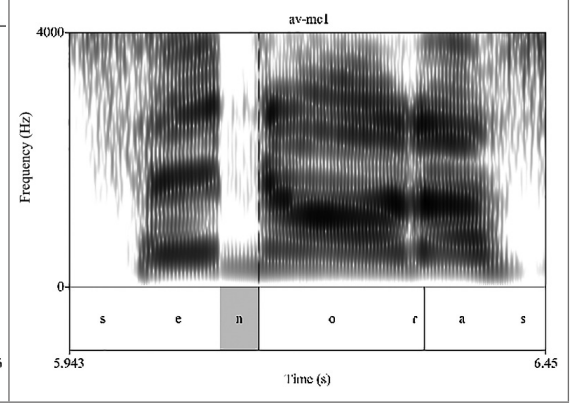
Tabla de rasgo (2)

Natural voice	***Artificial voice***
	

El tercer proceso (3) está ligado al segmento nasal palatal, al parecer FakeYou —en algunos contextos— genera una nasal alveolar en lugar de una nasal palatal.

Tabla 7

Tabla de rasgo (3)

Natural voice	***Artificial voice***
	

El cuarto proceso (4) se relaciona con la producción de la vibrante múltiple /r/. En la tabla 8, se aprecia que la variación idiolectal de la vibrante múltiple es bastante amplia en la muestra natural, mientras que en la muestra artificial la variación es mucho menor. De esta forma, en la muestra natural se ha reportado ocho variantes que se diferencian por el número de componentes, se ha reportado vibrantes múltiples desde los dos hasta los nueve componentes. Por el lado de la muestra artificial, se ha reportado únicamente tres variantes que van desde un componente hasta tres componentes. Además, se ha observado que la variante de un componente (aproximante) representa casi el 80 % de los ejemplos en la muestra artificial, mientras que en la muestra natural dicha variante no aparece y, asimismo, tampoco hay una variante que represente la mayoría de casos. En la muestra natural, se puede resaltar las variantes de cinco y siete componentes (5C y 7C) con un porcentaje de casos del 37 % y 26 % respectivamente, las otras seis variantes se distribuyen de forma más uniforme.

Tabla 8

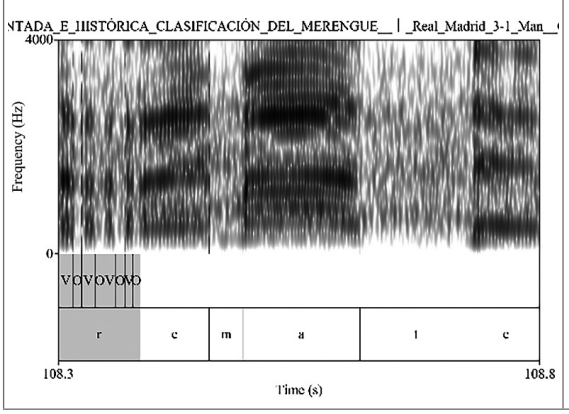
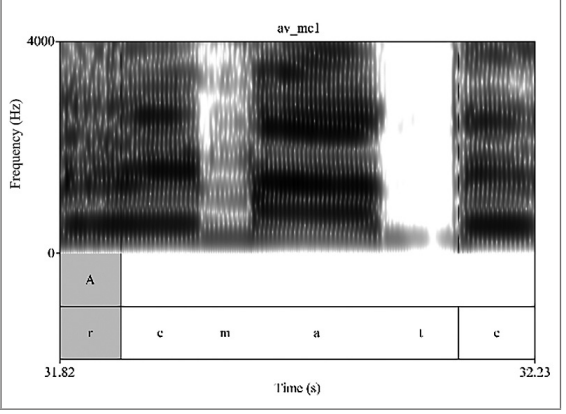
Distribución de la vibrante múltiple /r/ en las muestras

Natural voice			*Artificial voice*		
Número de componentes	**Frecuencia de aparición**	**Porcentaje**	**Número de componentes**	**Frecuencia de aparición**	**Porcentaje**
1C	0	0 %	1C	21	78 %
2C	1	4 %	2C	2	7 %
3C	2	7 %	3C	4	15 %
4C	2	7 %	Total	27	
5C	10	37 %			
6C	3	11 %			
7C	7	26 %			
8C	1	4 %			
9C	1	4 %			
Total	27				

En la tabla 9, se observa el contraste entre ambas muestras, en la muestra natural se reporta la variante múltiple de ocho componentes —entre elementos vocálicos (V) y oclusiones (O)—, mientras que en la muestra artificial se presenta la variante de un componente —de tipo aproximante—, esta última se caracteriza por poseer una estructura formántica similar a la de una vocal.

Tabla 9

Tabla de rasgo (4)

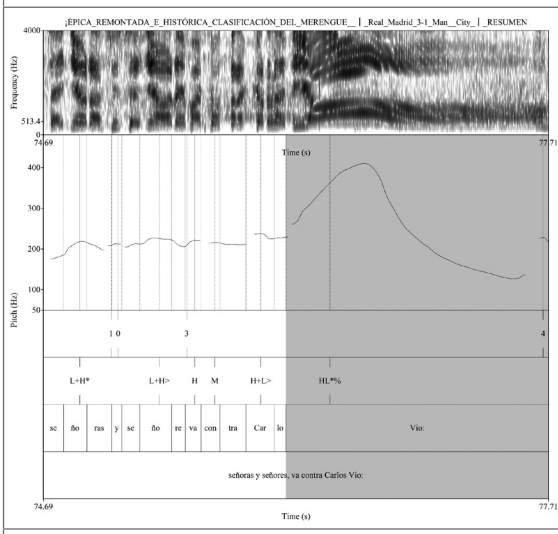
Natural voice	*Artificial voice*
	

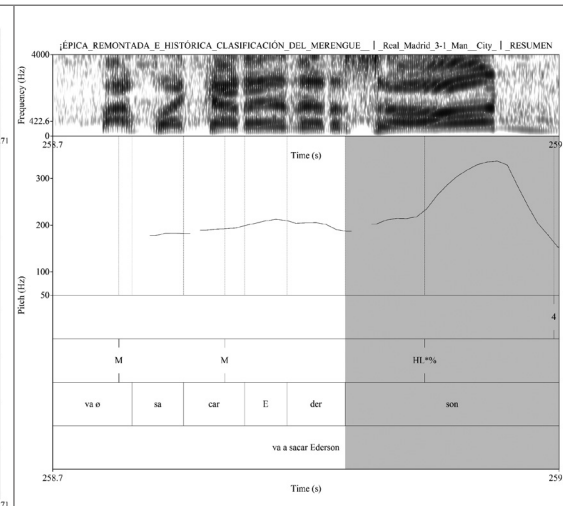
El quinto proceso (5) de la tabla 9 se relaciona con la entonación. Se puede apreciar que, en la voz natural, la entonación de la vocal alargada a final de palabra tiene una curvatura que asciende al inicio y desciende al final; mientras que en la muestra artificial, la entonación es descendente. Asimismo, se observa que el rango entonativo del Pitch es totalmente distinto en ambas muestras.

Tabla 10

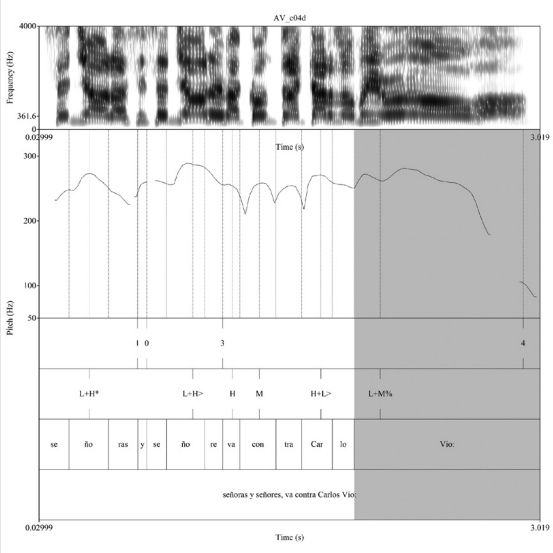
Tabla de rasgo (5)

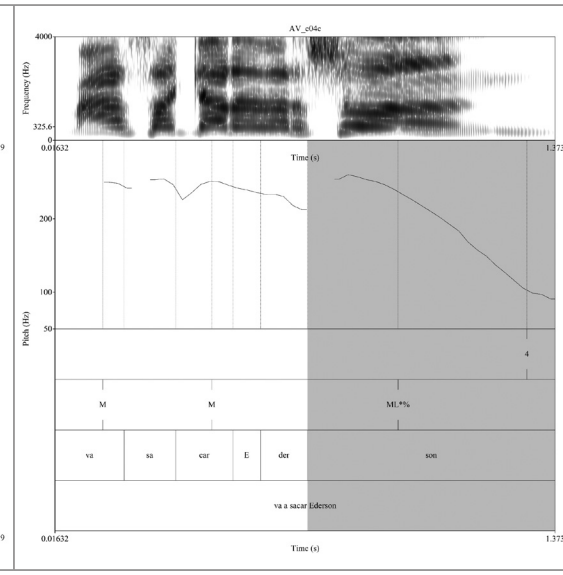
Natural voice





Artificial voice





Como se ha podido observar, se ha reportado procesos fonéticos (que se relacionan con algunas deficiencias del convertidor a texto FakeYou) que pueden ser sencillamente solucionados con una programación más minuciosa de la IA. Asimismo, se reportaron procesos más interesantes —para los fines del presente estudio—: el análisis de la vibrante múltiple /r/ mostró que la variación idiolectal es mucho más rica en la voz natural que en la voz artificial y el análisis de la entonación permitió encontrar patrones divergentes para las muestras.

5. CONCLUSIONES

Los métodos usados en el análisis de las muestras presentan resultados distintos. En primer lugar, los resultados de la comparación automática indican que la voz natural y la voz artificial de Mariano Closs (proveniente del sitio web FakeYou) pertenecen al mismo locutor, esto se aprecia al observar los valores del *likelihood ratio* (LR). No obstante, en SIS II, se pudo observar que mientras los valores eran altos para los parámetros de estructura formántica y GMM, eran bajos para el *pitch*. Puede haber distintos factores que afecten este parámetro, por lo que este resultado puede ser un indicio para hacer hincapié en dicho parámetro cuando se comparen muestras artificiales con muestras naturales. En segundo lugar, en el análisis fonético, se encontraron procesos que reflejaban desperfectos de la voz artificial de Mariano Closs (estos pueden ser subsanados con una programación más exhaustiva del convertidor de texto en habla). Asimismo, el análisis de la vibrante múltiple /r/ permitió observar que la variación idiolectal es más rica en la voz natural que en la voz artificial; por su lado, el análisis de la entonación mostró disimilitudes entre las muestras. Es así que se puede concluir que para el análisis de una voz artificial, el análisis fonético —y de forma más amplia el lingüístico— es sumamente importante porque da cuenta de características que pueden ayudar a diferenciar la voz natural de la voz artificial, a su vez que muestra la precisión con la que la inteligencia artificial replica los sonidos y los patrones suprasegmentales del hablante.

También es importante señalar que este estudio es de tipo exploratorio y que el análisis estuvo sujeto a las limitaciones que planteaba la naturaleza de las muestras de voz, por lo tanto, los resultados deben tomarse en cuenta con cautela. No obstante, la presente investigación significa un aporte al campo de la fonética forense porque plantea un antecedente metodológico para los especialistas que en algún momento pueden analizar casos como el planteado en este estudio. Además, el estudio también representa un aporte al campo judicial porque expone de forma clara (para profesionales del derecho, abogados y fiscales) la relevancia de la lingüística en la administración de justicia. Es así que se ha podido observar que la metodología empleada se lleva a cabo de manera rigurosa y sistemática, utilizando criterios técnicos con aceptación de la comunidad científica, lo que implica que la evidencia lingüística puede contribuir en el apoyo de la resolución de disputas judiciales.

Finalmente, a medida que el desarrollo de la IA avanza y se crean nuevos algoritmos y modelos de aprendizaje, es probable que las voces generadas artificialmente mejoren su calidad y similitud con las voces humanas. Por ende, la clonación de voz mediante inteligencia artificial plantea preocupaciones éticas y de seguridad; para mitigar estos riesgos, se debe establecer regulaciones, sistemas de autenticación de voz y marcas de agua auditivas. Asimismo, la conciencia pública y la educación son esenciales para prevenir la manipulación de información; por su lado, las empresas también deben ser responsables y transparentes para así promover un uso ético de los nuevos avances que se desarrollan en el marco de la inteligencia artificial.

REFERENCIAS

Boersma, P. y Weenink, D. (2023). *Praat: Doing Phonetics by Computer* (Versión 6.3.14) [Programa de computadora]. https://www.fon.hum.uva.nl/praat/

Bonafonte, A. (1997). Tecnologías del habla: conversión de texto a voz. *Buran,* (9), 68-72. https://core.ac.uk/reader/39120110

Cortez, A., Vega. H. y Pariona, J. (2009). Procesamiento del lenguaje natural. *Revista de Ingeniería de Sistemas e Informática*, *6*(2), 45-54. https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923/5121

Echelon (s. f.). FakeYou. Deep Fake Text to Speech. https://fakeyou.com/

ESPN Fans (2022a). *¡Benzema brilló y el merengue borró al PSG de Messi de la Champions!* | *Real Madrid 3-1 PSG* | *Resumen* [Video]. YouTube. https://www.youtube.com/watch?v=4jK2vjqcO5o

ESPN Fans (2022b). *¡Épica remontada e histórica clasificación del merengue!* | *Real Madrid 3-1 Man. City* | *Resumen* [Video]. YouTube. https://www.youtube.com/watch?v=lme15YYJUtQ

Fernández, A. M. (2007). ¿Para qué sirve la fonética? *Onomázen,* (15), 39-51. https://doi.org/10.7764/onomazein.15.02

Garayzábal, E., Queralt, S. y Reigosa, M. (2019). *Fundamentos de la lingüística forense*. Síntesis.

Gold, E. y French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech, Language and the Law*, *18*(2), 293-307. https://doi.org/10.1558/ijsll.v18i2.293

Jimenez, J., Torres, F. y Cueva, O. (2022). Identificación de locutor a partir de la fonética forense: aplicación del software SplitsTree4 para una organización esquemática de los datos lingüísticos. *Boletín de la Academia Peruana de la Lengua*, *71*(71), 431-461. https://doi.org/10.46744/bapl.202201.014

Lazo, V. (2023). La adecuación de la muestra indubitada en la comparación forense de voz. *Escritura y Pensamiento*, *22*(47), 179-205. https://revistasinvestigacion.unmsm.edu.pe/index.php/letras/article/view/25814/19896

Lazo, V. y Rivas, G. (2022) La relación entre el extorsionador y la víctima en un caso de extorsión: una aproximación desde el análisis de la conversación. *Lengua y Sociedad*, *21*(2), 373-400. https://revistasinvestigacion.unmsm.edu.pe/index.php/lenguaysociedad/article/view/22535/18891

Llisterri, J., Carbó, C., Machuca, M. J., Mota, C. de la, Riera, M. y Ríos, A. (2004). La conversión de texto en habla: aspectos lingüísticos. En M. Martí y J. Llisteri (eds.), *Tecnologías del texto y del habla* (pp. 145-186). Edicions de la Universitat de Barcelona – Fundación.

Machuca, M., Ríos, A. y Llisterri, J. (2014). Conocimiento fonético y fonética judicial. *Quaderns de Filología: Estudis Lingüístics, 19*, 95-111. https://ojs.uv.es/index.php/qfilologia/article/view/5188/4989

Mena, J. y Rojas, J. (2021). *Estado del arte del reconocimiento de voz artificial*. [Tesis para optar el título de ingeniero de sistemas y computación, Universidad Tecnológica de Pereira]. https://repositorio.utp.edu.co/server/api/core/bitstreams/a39928f4-b645-46a8-999d-54ba71ae00fd/content

Morrison, G. (2011). La comparación forense de la voz y el cambio de paradigma (C. Curiá, trad.). *Estudios Fónicos/Cuadernos de Trabajo,* (1), 1-38. (Obra original publicada en 2009)

Muñoz, R. (2020). *TgDraw* [Praat *plug-in*] (versión 0.3) [*Software*]. https://rolandomunoz.github.io/praat_tools/tg_draw.html

Olsson, J. (2008). *Forensic Linguistics* (2.ª ed.). Continuum.

Peña, J. (2022). Inteligencia artificial para la seguridad jurídica. Superando el problema de la cognoscibilidad del derecho. *Revista Oficial del Poder Judicial*, *14*(17), 55-117. https://revistas.pj.gob.pe/revista/index.php/ropj/article/view/568/754

Ramírez, S. (2023, 9 de septiembre). Clonan voces de personas con IA para estafar o fingir secuestros: al menos 55 casos en el Perú. *El Comercio*. https://elcomercio.pe/lima/clonacion-de-voz-para-estafar-con-inteligencia-artificial-como-funciona-esta-modalidad-y-que-recomendaciones-seguir-inseguridad-deepfake-ciberdelincuencia-hackers-secuestros-noticia/?ref=ecr

Rosas, C., Sommerhoff, J., Sáez, C. y Saavedra, S. (2011). Comparación de voz bajo el cociente de probabilidad en el caso de Luis Tralcal. *Revista de Lingüística Teórica y Aplicada*, *52*(1), 13-33. https://www.scielo.cl/pdf/rla/v52n1/art_02.pdf

San Segundo, E. (2022). *How deepfake is your voice? Understanding the linguistic foundations of deepfakes*. Github. https://eugeniasansegundo.github.io/project/deepfakes/

Speech Techonology Center. (2015). SIS II (versión 2.6.357) [Software Trial]. https://es.speechpro.com/product/analisis/ikarlab#tab3

Torres, F. (2023). *Identificación de locutor en el marco de la fonética forense en el Perú*. [Tesis de maestría]. Pontificia Universidad Católica del Perú.

Univaso, P. (2016). *Identificación forense de hablantes: un tutorial*. https://www.researchgate.net/publication/303639465_Univaso_Tutorial_Identificacion_Forense_de_Hablantes_2016_2

Univaso, P., Gurlekian, J., Martínez Soler, M. y Stalker, G. (2020). FORENSIA: un sistema de identificación forense por voz. Anales de SID 2020. *Simposio Argentino de Informática y Derecho (JAIIO)*, 116-130.

Financiamiento
Autofinanciado.

Conflicto de intereses
Los autores declaran no tener conflicto de intereses.

Contribución de autoría
Jhon Jimenez Peña: análisis e interpretación de datos, concepción y diseño del trabajo, redacción y revisión crítica; aprobación final de la versión que se publicará. Responsabilidad en la supervisión y el liderazgo para la planificación y la ejecución de la actividad de investigación, incluyendo las tutorías externas.

Fernando Aaron Torres Castillo: análisis e interpretación de datos, concepción y diseño del trabajo, redacción y revisión crítica; aprobación final de la versión que se publicará. Verificación, ya sea como parte de la actividad o por separado, de la replicación/reproducibilidad general de los resultados/experimentos y otros resultados de investigación.

Oscar Esaul Cueva Sanchez: análisis e interpretación de datos, concepción y diseño del trabajo, redacción y revisión crítica; aprobación final de la versión que se publicará. Preparación, creación y/o presentación del trabajo publicado, específicamente, la visualización/presentación de datos.

Agradecimientos
Agradecemos a los revisores de la *Revista Oficial del Poder Judicial* por sus comentarios sustanciales al estudio. También a Akuma por sus atinados comentarios al conversar sobre el tema.

Biografía de los autores

Jhon Jimenez Peña
Es licenciado en Lingüística por la Universidad Nacional Mayor de San Marcos (UNMSM). Sus intereses están centrados en la fonética y la fonología de las lenguas originarias del Perú, con especial atención a la lengua arabela. Ha sido consultor en el Ministerio de Educación para la elaboración de fonologías que se han empleado en los procesos de normalización de alfabetos del arabela, el ocaina y el taushiro. También ha sido docente de los cursos de Fonología y Fonología Avanzada en el Curso Internacional de Lingüística, Traducción y Alfabetización (CILTA) del Instituto Lingüístico de Verano en los años 2018 a 2023, que se imparte en la Universidad Ricardo Palma. Además, ha sido expositor para el primer «Curso-Taller de fonética forense» organizado por el CILA-UNMSM. Es miembro del grupo de investigación Dolenper: Documentación lingüística de lenguas amenazadas en el Perú (CILA-UNMSM). Actualmente, labora como perito lingüista forense en la Oficina de Peritajes del Ministerio Público-Fiscalía de la Nación y es miembro del Gabinete de Lingüística Forense del CILA-UNMSM.

Fernando Aaron Torres Castillo
Es licenciado en Lingüística por la Universidad Nacional Mayor de San Marcos (UNMSM), maestro en Lingüística por la Pontificia Universidad Católica del Perú (PUCP). Sus intereses giran en torno al estudio de lenguas amerindias, entre ellas las familias quechua y arawak. Actualmente labora como lingüista forense en la Oficina de Peritajes del Ministerio Público-Fiscalía de la Nación. Asimismo, es miembro adherente del grupo de investigación Kawsasun: Investigación intercultural para la formación docente y enseñanza de lenguas, del Instituto de Investigación de Lingüística Aplicada (CILA). También está adscrito como miembro del Gabinete de Lingüística Forense de la UNMSM.

Oscar Esaul Cueva Sanchez
Es licenciado en Lingüística por la Universidad Nacional Mayor de San Marcos (UNMSM). Sus intereses giran en torno a las áreas de fonética y fonología con especial atención al campo de la fonética acústica. Asimismo, es miembro del Gabinete de Lingüística Forense del Instituto de Investigación de Lingüística Aplicada (CILA).

Correspondencia
fernando.torresc@pucp.edu.pe