ALPHA
Nº62 2026
ISSN 0716-4254

https://doi.org/10.32735/S0718-22012026000624196 209-225

¿QUÉ TAN BIEN DISTINGUIMOS VOCES GENERADAS POR IA VERSUS VOCES GENERADAS POR HUMANOS? EL CONTEXTO DE LAS ESTAFAS TELEFÓNICAS

How well do we distinguish AI-generated voices versus human-generated voices? The context of phone scams

CLAUDIA ROSAS
Universidad Austral de Chile (Chile)
https://orcid.org/0000-0002-8544-7965
claudiarosas@uach.cl

MATEO CASTRO
Universidad Austral de Chile (Chile)
https://orcid.org/0009-0002-5671-702X
mateo.castro@alumnos.uach.cl

JORGE GUZMÁN
Policía de Investigaciones de Chile(Chile)
https://orcid.org/ 0009-0006-2602-6171
jguzmanm@investigaciones.cl

Resumen

La tecnología actual permite la generación de voz artificial. Existen numerosos softwares disponibles en la web de forma gratuita que permiten reproducir voces a partir de una sencilla muestra de un(a) hablante original. Estas aplicaciones ya se han ocupado para realizar estafas telefónicas. Evaluamos el desempeño de oyentes no expertos(as) para distinguir voces humanas familiares versus voces generadas por inteligencia artificial (IA) en condiciones de casos reales.

Palabras clave: Reconocimiento auditivo por oyentes no expertos(as); inteligencia artificial; clonación de voz.

Abstract

Current technology allows for artificial voice generation. There are numerous software programs available on the web for free that allow reproducing voices from a simple sample of an original speaker. These applications have already been used to conduct telephone scams. We evaluated the performance of non-expert listeners in distinguishing familiar human voices versus voices generated by artificial intelligence (AI) under real-life conditions.

Keywords: Auditory recognition by non-expert listeners; artificial intelligence (AI); voice cloning.

Recibido: 11 marzo 2025
Aceptado:11 abril 2025

1. INTRODUCCIÓN

En el presente artículo nos proponemos evaluar el desempeño de oyentes no expertos(as) para reconocer voces humanas familiares reales e identificar voces familiares falsas generadas con inteligencia artificial. Este objetivo se inscribe en el ámbito de la ciencia forense del habla y se relaciona, por una parte, con la habilidad humana natural de reconocer voces sin ningún tipo de entrenamiento, conocida como reconocimiento y/o identificación de voz, y, por otra parte, con el uso de tecnología para replicar digitalmente la voz de un(a) hablante a partir de una muestra de su propia voz, lo que se conoce comúnmente como clonación de voz.

En el ámbito forense, la clonación de voz aparece típicamente en las estafas telefónicas; por ejemplo, un delincuente extrae una muestra de audio de una adolescente desde YouTube y mediante un software obtiene nuevas grabaciones de su voz que usa para llamar a sus padres fingiendo un secuestro y exigirles dinero. En el presente artículo, intentamos reproducir, de manera general, las condiciones de este tipo de casos, incluyendo el perfil no experto del(la) oyente, las características de los audios, obtenidos de Internet, y la familiaridad de las voces escuchadas.

La detección de voces clonadas, generadas con inteligencia artificial, tiene un alto impacto negativo en la sociedad a nivel mundial; a pesar de ello, no hemos encontrado estudios que aborden específicamente el comportamiento perceptivo de oyentes comunes ante voces clonadas en contextos forenses que reflejan situaciones reales. Como parte de las condiciones de los casos, hay que considerar que, en la actualidad, la difusión del desarrollo tecnológico ha creado conciencia en la población respecto del potencial engaño que puede existir en una llamada telefónica. De manera que una pregunta de interés de investigadores(as), pero también del gobierno, ciudadanos(as) y fuerzas de orden y seguridad sería: ¿qué tan bien oyentes no expertos(as) distinguen voces familiares generadas por IA *versus* voces familiares generadas por humanos en el contexto potencial de una llamada fraudulenta? La respuesta a esta pregunta contribuirá a esclarecer de manera más consecuente la real magnitud de impacto de las voces clonadas.

1.1. RECONOCIMIENTO DE VOCES FAMILIARES HUMANAS REALES E IDENTIFICACIÓN DE VOCES FAMILIARES ARTIFICIALES

Existe una cantidad importante de literatura de investigación sobre el reconocimiento de hablantes y la identificación de hablantes por oyentes inexpertos(as). Revisiones de la literatura de investigación desde la perspectiva de la posible aplicación en contextos legales se encuentran en Rose, 2002; Solan y Tiersma, 2003; Yarmey et al., 2007; Sherrin, 2016; y Morrison et al., 2018; y estudios centrados en cuestiones legales relacionadas en Ormerod (2001); Edmond y San Roque (2009); Edmond et al. (2011); Laub et al. (2013); Robson (2018). En dicha literatura, "reconocimiento de hablante" se

refiere a la situación en la que un(a) oyente escucha una voz (en vivo o grabada) y afirma que reconoce la voz como la de una persona que le resulta familiar (y normalmente nombra a esa persona). Este escenario contrasta con la "identificación del(la) hablante" que se refiere a una situación en la que un(a) oyente que no está familiarizado(a) con el(la) hablante o los(as) hablantes compara una voz que escucha en una ocasión (por ejemplo, mientras se está cometiendo un crimen) con una voz que escucha en otra ocasión (por ejemplo, durante una rueda de reconocimiento de voces) y, basándose en la escucha, intenta determinar si el(la) mismo(a) hablante estaba hablando en ambas ocasiones. "Identificación del(la) hablante" también se refiere a una situación en la que un(a) oyente que no está familiarizado(a) con el(la) hablante o los(as) hablantes escucha dos (o más) grabaciones de voz y, basándose en la escucha, intenta determinar si el(la) mismo(a) hablante está hablando en ambas grabaciones (Rosas et al., 2019; Bali et al., 2024). El presente artículo se enfoca en el reconocimiento del(la) hablante, no en la identificación del(la) hablante.

Por su parte, el término "familiar" en la literatura se usa generalmente para describir una voz que el(la) oyente ha escuchado en muchas ocasiones durante largos períodos de tiempo (usualmente, años) unido a una larga duración (al menos, muchas horas) de exposición a la voz. Típicamente, incluiría interacciones del tipo padres e hijos, pero también compañeros(as) de trabajo, profesores(as) y estudiantes, médicos y pacientes e incluso, aunque en menor grado, personajes famosos y teleaudientes. Estas relaciones implican una exposición sustancial a la variabilidad de la voz dentro del(la) hablante en diferentes contextos (diferentes grados de formalidad, emocionalidad, ambientes, etc.). Dada la cantidad de factores que intervienen en su construcción, la "familiaridad" no es una cuestión binaria de presencia *vs* ausencia, sino una situación en la que los(as) oyentes pueden estar más familiarizados(as) con algunas voces y menos familiarizados(as) con otras (Yarmey et al., 2001).

Otro sentido en el que el término "familiar" se ha utilizado en entornos forenses se aplica en la situación en la que un(a) oyente escucha repetidamente grabaciones de audio de la voz de un(a) hablante para "familiarizarse" deliberadamente con la voz. Se trataría, por ejemplo, de escenarios de experimentos auditivos en condiciones controladas. En estas situaciones es poco probable que el procedimiento empleado exponga al (a la) oyente a la misma duración y variedad de la voz del(la) hablante durante el mismo período de tiempo que sería el caso para familiares y amigos, e incluso personalidades de los medios. Debido a esto, se esperaría que el rendimiento de reconocimiento de los(as) oyentes en los(as) hablantes "familiarizados" (en laboratorios) sea más pobre que en los(as) hablantes muy familiares, como parientes cercanos(as) y amigos(as) (Morrison et al., 2018). En el presente artículo, usamos "familiar" en el primer sentido descrito más arriba.

1.2. Inteligencia artificial (IA) y clonación de voces.

Existen numerosas publicaciones sobre inteligencia artificial. Revisiones desde la perspectiva forense pueden encontrarse en Moustafa (2022); Ypma et al. (2023); Saini et al. (2024); sobre los fundamentos de la inteligencia artificial y sus diversas aplicaciones, en LeCun et al. (2015); Neapolitan et al. (2018); Saguna et al. (2021); y estudios específicos que abordan la clonación de voz, incluyendo estrategias de detección, en Mcuba et al. (2023).

La inteligencia artificial se basa en algoritmos complejos, los métodos que los utilizan y sistemas, entendidos estos como la(s) herramienta(s) de software que implementan ese método y algoritmo (Ypma et al. (2023). La clonación de voz es una aplicación, basada en tecnología de redes neuronales profundas que forman el llamado campo de aprendizaje profundo (*deep learning*). Una red neuronal profunda (DNN) es una red densa y muy compleja de conexiones entre unidades llamadas neuronas. Estas conexiones pueden incluso ser recurrentes, o complementadas con filtros que interrelacionan diferentes regiones en las características de entrada. Originalmente, una red neuronal está destinada a imitar la topología del cerebro humano, de ahí su nombre (LeCun et al., 2015). En el caso específico de la clonación de voz, estas neuronas ordenadas en capas se especializan en extraer datos específicos, yendo desde las frecuencias fundamentales y formantes, y pasando por frecuencias cepstrales de mel (MFCC), hasta llegar a representar perfectamente los sonidos del habla. Lo sorprendente es cómo a partir de breves emisiones de voz, estas neuronas capturan patrones de habla como respiración, ritmo y entonación, además de consonantes y vocales y crean una réplica digital de la voz de una persona (Napolitano, 2020; Arik et al., 2018; Zhao & Chen, 2020).

Los métodos de clonación han traído beneficios a la sociedad mediante la entretención, la educación, y el comercio, pero también riesgos cuando son utilizados maliciosamente para realizar estafas. Los ciberdelincuentes utilizan la tecnología de clonación de voz para hacerse pasar por celebridades, autoridades o personas comunes y corrientes, con fines fraudulentos. El problema que surge es que los audios falsos son difíciles de detectar por humanos (Saleema y Thampi, 2018) y, todavía más, cuando se acompañan de video. Un ejemplo reciente que ilustra esto se observó en la reciente guerra entre Rusia y Ucrania, donde hubo un video falso, que se viralizó, del presidente Zelenskyy diciendo a los soldados que entregaran las armas y se rindieran (Simonite, 2022).

Pese a su perjudicial impacto no se han encontrado estudios que aborden la percepción de la clonación en condiciones realistas, con excepción de Castro (2023). Este autor evaluó la capacidad discriminante de oyentes expertos(as) y no expertos(as), sin embargo, no se enfocó en las condiciones reales de audios que incluyen voces familiares. Así, la detección de audios falsos plantea un reto emergente a la ciencia forense, específicamente a una de sus ramas, la ciencia forense digital o multimedia, que se ocupa

de investigar la autenticidad (o no) de un archivo multimedia determinado (Zawali et al., 2021), y abre un campo que recién comienza a explorarse (Mcuba et al., 2023).

En el presente artículo nuestro foco son los(as) oyentes no expertos(as) de los audios clonados, no la evaluación forense, ni la detección de las propiedades acústicas que inducen al engaño.

2. Método

2.1. Hablantes

Los(as) hablantes consistieron en 4 políticos(as) chilenos(as) famosos(as): Michelle Bachelet, Evelyn Matthei, Sebastián Piñera y Gabriel Boric. En adelante, usaremos solo el apellido para referirnos a ellos(as).

2.2. Estímulos

Se buscaron en Internet grabaciones o entrevistas de los hablantes seleccionados con una mínima cantidad de ruido y consistentes en su acento y vocalización, la base de datos tenía una duración de alrededor de 20 minutos por cada uno.

Las secciones se seleccionaron manualmente de la grabación de cada hablante, con la condición de que contuvieran solo el discurso del (la) hablante de interés.

Luego se extrajeron de la grabación de cada hablante tres secciones de una duración de ~15 s.

Finalmente, para cada sección de habla real seleccionada se generó una sección nueva con inteligencia artificial (en adelante, IA) que pareciera idéntica a la original, que llamaremos "clonada" o "IA".

El número total de estímulos fue de 24 (seis secciones, 3 reales y 3 IA x 4 hablantes). Atendiendo experiencias anteriores, consideramos que un número discreto de estímulos beneficiaría la amplitud de participantes (Rosas et al., 2019).

Para la clonación se utilizó la aplicación web ElevenLabs (elevenlabs.io), la cual tiene una versión gratuita, pero en esta ocasión se compró la versión de pago más básica. Dicha aplicación desarrolla modelos de audio con IA que generan voces, efectos de sonido y discursos realistas, versátiles y contextualizados en +32 idiomas.

2.3. Oyentes

Los(as) oyentes fueron 62 personas adultas chilenas de entre 18 y 65 años. Se pidió a los(as) potenciales participantes que no colaboraran si tenían problemas de audición.

2.4. EXPERIMENTO DE AUDICIÓN

El experimento de escucha se presentó en línea a través de un navegador web. Los(as) oyentes vieron primero información relacionada con el consentimiento informado. Si aceptaban continuar, luego veían las instrucciones.

Se les preguntó para cada hablante famoso(a) si reconocerían su voz y si podrían detectar cuando su voz había sido clonada, solo si estas condiciones se cumplían podían continuar. En la pantalla, a la izquierda aparecía un botón de reproducción y a la derecha una escala de evaluación con la siguiente instrucción: "Seleccione un valor entre 1 y 10, que represente si la voz coincide con un hablante real o no en cada caso, donde 1 es muy real y 10 muy falso".

Al finalizar la prueba, aparecía en pantalla un botón de "enviar" que los(as) oyentes debía presionar; solo entonces sus respuestas serían consideradas en el análisis.

Antes del experimento propiamente tal, se solicitó a los(as) oyentes completar información demográfica general y se les preguntó si tenían algún conocimiento sobre clonación de voces.

2.5. CODIFICACIÓN DE DATOS

Los datos brutos consistieron en las respuestas escritas de cada oyente a cada sección de la grabación de cada hablante famoso(a).

Los datos fueron codificados y anonimizados. A cada hablante (político(a) famoso(a)) y oyente se le asignó un código numérico único, que se utilizó en lugar del nombre del(la) hablante y del(la) oyente.

La versión anonimizada de los datos se utilizó para todos los análisis posteriores.

2.6. ANÁLISIS ESTADÍSTICO

Se aplicó un enfoque tradicional que considera el conteo bruto de aciertos y desaciertos, independientemente del grado en que las respuestas ser acercan o alejan del valor verdadero, y un enfoque moderno que recoge las diferencias en términos de razones de verosimilitud (*LR*, sigla del inglés *Likelihood ratio*).

El grado en que el valor de una respuesta se acercaba o alejaba del valor verdadero estuvo dada por la diferente ubicación en la escala de 1 a 10 usada para calificar las muestras escuchadas, donde 1 era muy real y 10 muy falso. Para ilustrar esto: una respuesta de 1 para una muestra de voz real es más exacta que una respuesta de 5 para la misma muestra, pese a que ambas son correctas; mutatis *mutandi*s para las muestras clonadas, donde una respuesta de 10 es más exacta que una de 6.

La combinación de estos dos enfoques de análisis nos permitirá obtener una visión panorámica general y, también, más precisa de los resultados.

a) Matriz de las respuestas los(as) oyentes por hablante

Se confeccionó una matriz con los valores de las respuestas de cada oyente para cada muestra de voz de cada hablante.

b) Métricas para los resultados de desempeño de razones de verosimilitud (lr)

Las respuestas de cada oyente individual fueron tratadas como razones de verosimilitud (LR) para las cuales se calculó una métrica de desempeño que es estándar en sistemas de evaluación forense que dan resultados de razones de verosimilitud (Bali et al. 2024): C_{llr} (costo del logaritmo de la razón de verosimilitud).

El C_{llr} se usa para medir la exactitud de las respuestas; es decir, en nuestro caso, qué tan cerca de 1 se encuentra la respuesta de un(a) oyente si la voz era real (R) y qué cerca de10 se encuentra la respuesta de un(a) oyente si la muestra de voz era clonada (IA).

El procedimiento para obtener la LR y el C_{llr} fue el siguiente: Los datos de respuesta de cada oyente se dividieron entre audios R y audios IA. Cada respuesta daba 2 valores: uno a la hipótesis correcta y otro a la incorrecta, la hipótesis correcta e incorrecta varía según si el audio es real o IA. Para la razón de verosimilitud se calculó en base a la siguiente ecuación:

$$LR = \frac{p(E|H_r, I)}{p(E|H_f, I)} \quad (1)$$

Donde p representa probabilidad, Hr es la hipótesis del hablante correcto, Hf es la hipótesis del hablante incorrecto, I corresponde a la información dada sobre los audios y E corresponde a la evidencia presentada (Hughes y Rhodes, 2018).

	LR Hipótesis Correcta Hr (Same Speaker)	**LR Hipótesis Incorrecta Hf (Different Speaker)**
Audio Real	El audio es real (1 – 10)	El audio es falso (10 - 1)
Audio IA	El audio es falso (1 - 10)	El audio es real (10 – 1)

Por ejemplo:

- Si la respuesta del oyente cuando el audio es Real es 1, corresponde a un Hr = 10 y Hf = 1, en probabilidad 1 y 0.1, LR = 1/0.1 = 10, log10(LR) = 1.
- Si la respuesta del oyente cuando el audio es IA es 1, corresponde a un Hr = 10 y Hf = 1, en probabilidad 1 y 0.1, LR = 1/0.1 = 10, log10(LR) = 1.
- Si la respuesta del oyente cuando el audio es Real es 10, corresponde a un Hr = 1 y Hf = 10, en probabilidad 0.1 y 1, LR = 0.1/1 = 0.1, log10(LR) = -1.
- Si la respuesta del oyente cuando el audio es IA es 10, corresponde a un Hr = 1 y Hf = 10, en probabilidad 0.1 y 1, LR = 0.1/1 = 0.1, log10(LR) = -1.

En la Tabla 1, se presenta la escala con la que se convirtieron todos los valores de respuesta a fuerza de evidencia, probabilidades y LR:

Tabla 1 Conversión de los valores de respuesta a fuerza de evidencia, probabilidades y LR.

Audio Real								
Respuesta	Hr	Hf	P(Hr)	P(Hf)	LRss	LRds	log (LRss)	log (LRds)
1	10	1	1	0,1	10,0	0,1	1,00	-1,00
2	9	2	0,9	0,2	4,5	0,2	0,65	-0,65
3	8	3	0,8	0,3	2,7	0,4	0,43	-0,43
4	7	4	0,7	0,4	1,8	0,6	0,24	-0,24
5	6	5	0,6	0,5	1,2	0,8	0,08	-0,08
6	5	6	0,5	0,6	0,8	1,2	-0,08	0,08
7	4	7	0,4	0,7	0,6	1,8	-0,24	0,24
8	3	8	0,3	0,8	0,4	2,7	-0,43	0,43
9	2	9	0,2	0,9	0,2	4,5	-0,65	0,65
10	1	10	0,1	1	0,1	10,0	-1,00	1,00

Audio IA								
Respuesta	Hr	Hf	P(Hr)	P(Hf)	LRss	LRds	log (LRss)	log (LRds)
1	1	10	0,1	1	0,1	10,0	-1,00	1,00
2	2	9	0,2	0,9	0,2	4,5	-0,65	0,65
3	3	8	0,3	0,8	0,4	2,7	-0,43	0,43
4	4	7	0,4	0,7	0,6	1,8	-0,24	0,24
5	5	6	0,5	0,6	0,8	1,2	-0,08	0,08
6	6	5	0,6	0,5	1,2	0,8	0,08	-0,08
7	7	4	0,7	0,4	1,8	0,6	0,24	-0,24
8	8	3	0,8	0,3	2,7	0,4	0,43	-0,43
9	9	2	0,9	0,2	4,5	0,2	0,65	-0,65
10	10	1	1	0,1	10,0	0,1	1,00	-1,00

3. RESULTADOS

3.1. MATRIZ DE LAS RESPUESTAS DE LOS(AS) OYENTES POR HABLANTE

La Tabla 2 proporciona los recuentos brutos de la cantidad de veces que cada oyente respondió en una escala de 1 a 10 a cada estímulo, donde 1 representa "muy real" y 10, "muy falso". En cada celda el número desplazado hacia la izquierda de la celda y a la izquierda de la raya (/n) representa los aciertos, es decir, cuando el (la) oyente respondió que la muestra de la voz era real y era en verdad real (R) o clonada y era en verdad clonada (IA).

El número desplazado hacia la derecha de la celda y a la derecha de la raya (n/) representa, en tanto, cuando el(la) oyente respondió que la muestra era real y no era en verdad real (R) o clonada y no era en verdad clonada (IA). Los ceros representan los casos en que no existió respuesta.

De un número próximo a 100 participantes, solo 43 cumplieron con la condición establecida de declarar que podrían detectar la voz clonada (IA) de al menos un(a) hablante de interés. Aunque la confianza no está directamente relacionada con el grado de aciertos necesariamente (Morrison et al., 2018), se consideró relevante incluirla como

garantía de que quienes participaban, estaban realmente familiarizados(as) con las voces de la prueba.

De 1.032 oportunidades para dar una respuesta (6 muestras × 4 hablantes × 43 oyentes), hubo 822 respuestas en total. Esto se debió a que 35 oyentes declararon no reconocer la voz de un(a) hablante en particular o reconocer la voz, pero no ser capaz de detectar si había sido clonada (IA).

En cuanto a los(as) hablantes, la voz que presentó menos respuestas por oyentes fue de Matthei (16 oyentes no emitieron respuestas), sin embargo, obtuvo el más alto reconocimiento. La voz que presentó más respuestas por oyentes fue de Piñera (solo 3 oyentes no emitieron respuestas).

La observación más obvia de los resultados es que la mayoría de las combinaciones de oyentes y hablantes produjeron respuestas correctas. El reconocimiento para las voces reales (R) fue: 90,7% para la voz de Matthei; 89,5 %para la voz de Piñera, 87,7% para la voz de Boric; y 80,5% para la voz de Bachelet. Lo más sorprendente fue el alto porcentaje de reconocimiento de voces clonadas (IA): 84, 3% para la voz de Matthei; 75, 6% para la voz de Piñera; 60,1% para la voz de Bachelet; y 58,1% para la voz de Boric.

No tenemos una explicación de por qué las muestras clonadas (IA) fueron reconocidas correctamente a una tasa más alta, ni tampoco de por qué la voz clonada de Matthei obtuvo el mayor reconocimiento, pero sí podemos proponer como posible explicación para la variación interna entre los(as) 4 hablantes que la voz de Matthei era más familiar para quienes respondieron y eso les permitió detectar sutiles diferencias presentes en las muestras clonadas (IA) con respecto a las voces reales (R).

En cuanto a los(as) oyentes, de 43 participantes, solo 8 respondieron todas las secciones: 16 oyentes no respondieron la sección de Matthei, 11 oyentes, no respondieron la sección de Bachelet, 5 oyentes no respondieron la sección de Boric y 3 oyentes no respondieron la sección de Piñera. El(la) oyente 10 fue quien tuvo el mejor desempeño: respondió todas las secciones (24 respuestas) y obtuvo un rendimiento de 90,8% para las muestras de voces reales (R) y 98,3% para las muestras de voces clonadas (R).

Llama la atención que el rendimiento para las voces clonadas (IA) sea mejor que el de las voces reales (R), pero no es el único caso: el(la) oyente 30, igualmente respondió todas las secciones y obtuvo un 100% de reconocimiento para las voces clonadas (IA) y solo un 70% para las voces reales (R); los(as) oyentes(s) 22 y 31 de 12 respuestas, obtuvieron un 98,3% de reconocimiento para las muestras reales (R) y un 100% para las muestras clonadas (IA). Tampoco tenemos una explicación para este resultado en particular; podría estar asociado a una tendencia inconsciente (sesgo) que favorece la idea de que las voces escuchadas, dado el contexto en que se inscribe la prueba, son clonadas. De cualquier forma, su comprobación requiere ampliar la cantidad de muestras por hablante.

Tabla 2 Matriz de las respuestas de los(as) oyentes por hablante.

Hablantes	Piñera						Bachelet						Matthei						Boric											
ID de respuesta	R	R	R	IA	IA	IA	R	R	R	IA	IA	IA	R	R	R	IA	IA	IA	R	R	R	IA	IA	IA	# RESP	R	IA	V MAX	RESP T	REND	V MAX	R	IA
1	5/		/10	/1	10/	10/	1/		/8	/1	10/		2/	1/		2/	10/		/1	10/	5/		1/		/10	7/	/1	8/	24	77	71	240	148	61,7	120	64,2	59,2
2	1/		1/	2/	7/	8/	1/		/6	/3	7/		0	0		0	0		0	0	1/		1/		1/		/4	/4	18	84	46	180	130	72,2	90	93,3	51,11
3	/10	1/	1/	/2	10/	10/	1/		1/	/1	/1	10/	0	0		0	0		0	0	0		0		0	0	0	0	12	51	34	120	85	70,8	60	85,0	56,7
4	3/		1/	1/	/1	/1	10/		/7	3/	/2	/1	10/	1/		1/	10/	10/	1/	1/	1/	10/	10/	10/	24	110	85	240	195	81,3	120	91,7	70,8
5	1/		1/	6/		/1	10/		0	0	0	0	0	0		0	0	0	0	1/	1/	1/	10/	10/	12	60	47	120	107	89,2	60	100,0	78,3
6	1/		2/	1/	/1	9/	10/		/10	/10	10/	9/	10/	0		0	0	0	0	5/	1/	2/	1/	10/	18	67	70	180	137	76,1	90	74,4	77,8
7	1/		1/	2/	8/	9/	10/		1/	1/	/5	/5	/1	1/		1/	10/	10/	10/	10/	1/	5/		/3	24	115	81	240	196	81,7	120	95,8	67,5
8	1/		1/	2/	8/	9/	10/		1/	2/	2/	7/	8/	9/		1/	/6	5/	9/	9/	8/	1/	1/	1/	/2	/5	/1	24	108	85	240	193	80,4	120	90,0	70,8
9	1/		3/	1/	/2	9/	10/		0	0	0	0	0	2/		4/	3/	10/	10/	8/	1/	2/		/7	9/	10/	9/	18	75	77	180	152	84,4	90	83,3	85,6
10	1/		1/	9/	10/	10/	1/		1/	/8	10/	10/	10/	1/		5/	1/	10/	10/	10/	1/	1/	1/	10/	9/	10/		24	109	118	240	227	94,6	120	90,8	98,3
11	1/		1/	1/	9/	9/	1/		1/	/10	/1	9/	9/	0		0	0	0	0	0	1/	1/	1/	10/	/1	10/		18	81	67	180	148	82,2	90	90,0	74,4
12	1/		1/	1/	/2	/1	8/	7/	1/	1/	/1	/1	/12/	5/		1/	/1	9/	/1	14/	/8	/7	/1	/2	6/			24	99	34	240	133	55,4	120	82,5	28,3
13	2/		1/	1/	8/	8/	7/		0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0			6	29	23	60	52	86,7	30	96,7	76,7
14	2/		2/	1/	9/	8/	10/		1/	2/	2/	6/	/2	8/		2/	3/	1/	10/	10/	10/	3/	3/	5/	8/	9/	6/		24	105	96	240	201	83,8	120	87,5	80,0
15	1/		1/	1/	8/	10/	10/		1/	1/	5/	9/	/3	7/		2/	3/	1/	10/	9/	9/	1/	1/	1/	/2	/5	6/		24	113	88	240	201	83,8	120	94,2	73,3
16	1/		1/	1/	10/	10/	10/		1/	/6	7/	8/	6/	1/		1/	1/	10/	10/	10/	1/	1/	1/	/10	/1	8/	/1		24	106	91	240	197	82,1	120	88,3	75,8
17	1/		1/	1/	/1	10/	10/		0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0		6	30	21	60	51	85,0	30	100,0	70,0
18		0	0	0	0	0	0		0	0	0	0	0	0		0	0	0	0	0	1/	1/	3/	/3	/3	/2		6	28	8	60	36	60,0	30	93,3	26,7
19	1/		1/	/10	10/	/2	10/		1/	1/	/10	/1	10/	1/		1/	10/	10/	10/	10/	1/	1/	1/	10/	10/	10/	10/		24	102	94	240	196	81,7	120	85,0	78,3
20	1/		1/	/10	10/	10/	10/		1/	1/	10/	/1	/1	0		0	0	0	0	0	1/	1/	1/	10/	10/	10/	10/		18	81	76	180	157	87,2	90	90,0	84,4
21	1/		1/	1/	/1	10/	10/		1/	/10	/2	/5	10/	0		0	0	0	0	0	1/	1/	1/	/1	/5	10/		18	81	54	180	135	75,0	90	90,0	60,0
22		0	0	0	0	0	0		0	0	0	0	0	0		0	0	0	0	0	1/	1/	1/	10/	10/	10/		12	59	60	120	119	99,2	60	98,3	100,0
23	1/		3/	1/	8/	8/	10/		0	0	0	0	0	0		0	0	0	0	0	1/	1/	1/	/1	10/	8/		12	58	45	120	103	85,8	60	96,7	75,0
24	1/		2/	2/	9/	7/	9/		0	0	0	0	0	0		0	0	0	0	0	2/	2/	2/	/5	9/	/5		12	55	44	120	99	82,5	60	91,7	73,3
25	2/		2/	5/	9/	9/	6/		2/	1/	3/	9/	7/	7/		0	0	0	0	0	0	0	0	0	0	0	0		12	51	47	120	98	81,7	60	85,0	78,3
26	2/		2/	2/	9/	7/	6/		1/	1/	1/	4/	/1	/2		0	0	0	0	0	2/	1/	2/	1/	/1	/1	/1		18	85	39	180	124	68,9	90	94,4	43,3
27		0	0	0	0	0	0		0	0	0	0	0	0		0	0	0	0	0	1/	8/	/1	/6	1/	/1	/1		12	50	13	120	63	52,5	60	83,3	21,7
28	3/		2/	2/	10/	6/	/5	1/	1/	5/	8/	/5	/5	1/		1/	9/	10/	/1	1/	1/	2/	1/	1/	/1	/5	/1		24	107	66	240	173	72,1	120	89,2	55,0
29	1/		3/	4/	7/	9/	3/	1/	1/	/6	/2	/4	/4	1/		5/	2/	9/	10/	8/	4/	1/	4/	/1	/3	/5		24	97	69	240	166	69,2	120	80,8	57,5
30	/10		/10	1/	10/	10/	10/	1/	/10	1/	10/	10/	10/	1/		1/	1/	10/	10/	10/	1/	1/	1/	/10	10/	10/		24	84	120	240	204	85,0	120	70,0	100,0
31	1/		1/	2/	10/	10/	10/	0	0	0	0	0	0	1/		1/	1/	10/	10/	10/	1/	1/	1/	0	0	0		12	59	60	120	119	99,2	60	98,3	100,0
32	1/		1/	1/	10/	8/	1/	1/	/8	10/	10/	10/	/10	2/		1/	10/	10/	10/	9/	1/	1/	1/	8/	/2	/2	/2		24	103	103	240	206	85,8	120	85,8	85,8
33	1/		1/	2/	9/	8/	2/	2/	/8	7/	7/	/2	/2	2/		2/	9/	8/	9/	9/	2/	2/	2/	1/	/2	/2	/2		24	104	67	240	171	71,3	120	86,7	55,8
34	1/		1/	1/	10/	10/	10/	1/	1/	/10	10/	10/	/1	1/		/10	1/	10/	10/	10/	10/	1/	1/	1/	/1	10/	10/		24	102	102	240	204	85,0	120	85,0	85,0
35	/7		1/	1/	10/	10/	10/	1/	/10	8/	8/	10/	1/	1/		1/	10/	10/	10/	9/	1/	1/	1/	1/	/1	/2	/2		24	105	90	240	195	81,3	120	87,5	75,0
36	1/		1/	1/	7/	/5	/4	1/	1/	2/	10/	7/	/1	1/		1/	1/	6/	/5	/4	1/	1/	1/	8/	/5	/1		24	119	63	240	182	75,8	120	99,2	52,5
37	0		0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	/10	/10	1/	/3	10/	10/		6	12	23	60	35	58,3	30	40,0	76,7
38	1/		1/	1/	/1	/18/	1/	1/	1/	1/	/1	/1	/1	0		0	0	0	0	0	1/	1/	1/	1/	/1	/1		18	90	16	180	106	58,9	90	100,0	17,8
39	1/		1/	1/	9/	10/	1/	1/	2/	6/	9/		/4	1/		1/	10/	10/	10/	10/	1/	1/	1/	10/	9/	7/		24	119	104	240	223	92,9	120	99,2	86,7
40	1/		1/	1/	/2	10/		/1	3/	4/		/10	10/	10/	1/	1/		/3	10/	10/	1/	2/		/10	10/	10/	10/		24	96	96	240	192	80,0	120	80,0	80,0
41	/6		6/	5/	6/	7/	7/	4/	5/	/6	6/	7/	6/	5/		4/	4/	7/	7/	6/	4/	4/	5/	/5	7/	6/		24	74	77	240	151	62,9	120	61,7	64,2
42	2/		1/	2/	7/	9/	8/	1/	1/	5/	9/	7/	7/	2/		1/	2/	10/	9/	10/	1/	1/	2/	6/	/5	6/		24	111	93	240	204	85,0	120	92,5	77,5
43	1/		1/	1/	/5	/4	6/	/8	4/	4/	/4	/2	6/	1/		1/	1/	/4	/4	/3	1/	1/	4/	1/	/3	/2		24	104	44	240	148	61,7	120	86,7	36,7
PUNTUACIÓN	348	355	344	247	301	337	304	283	186	190	186	201	250	228	257	235	232	216	344	349	307	179	245	238								
# RESP	39	39	39	39	39	39	32	32	32	32	32	32	27	27	27	27	27	27	38	38	38	38	38	38								
V MAX	390	390	390	390	390	390	320	320	320	320	320	320	270	270	270	270	270	270	380	380	380	380	380	380								
RENDIMIENTO%	89,2	91,0	88,2	63,3	77,2	86,4	95,0	88,4	58,1	59,4	58,1	62,8	92,6	84,4	95,2	87,0	85,9	80,0	90,5	91,8	80,8	47,1	64,5	62,6								
PUNTUACIÓN	1047,0			885,0			773,0			577,0			735,0			683,0			1000,0			662,0										
V MAX	1170,0			960,0			810,0			1140,0																						
RENDIMIENTO%	89,5			75,6			80,5			60,1			90,7			84,3			87,7			58,1										
PUNTUACIÓN	1932						1350						1418						1662													
V MAX	2340,0						1920,0						1620,0						2280,0													
RENDIMIENTO%	82,6						70,3						87,5						72,9													

3.2 MÉTRICAS PARA RAZONES DE VEROSIMILITUD (LR)

Se calculó un valor de C_{llr} por separado para las respuestas de cada oyente individual. Los valores menores a 1 indican que el resultado proporciona información útil (Bali et al., 2024, Morrison et al. (2021).

Hubo variabilidad entre oyentes, sin embargo, el desempeño general de todos(as) los(as) oyentes fue de 0, 7. Esta información se correlaciona con la interpretación general de la matriz de oyentes x audios. La Figura 1 muestra un gráfico de Tippett con el resultado del desempeño general. La línea azul representa los valores de las respuestas cuando los(as) oyentes proporcionan una respuesta que corresponde con la realidad y la línea roja, cuando no coincide con la realidad.

3.2 MÉTRICAS PARA RAZONES DE VEROSIMILITUD (LR)

Se calculó un valor de C_{llr} por separado para las respuestas de cada oyente individual. Los valores menores a 1 indican que el resultado proporciona información útil (Bali et al., 2024, Morrison et al. (2021).

Hubo variabilidad entre oyentes, sin embargo, el desempeño general de todos(as) los(as) oyentes fue de 0, 7. Esta información se correlaciona con la interpretación general de la matriz de oyentes x audios. La Figura 1 muestra un gráfico de Tippett con el resultado del desempeño general. La línea azul representa los valores de las respuestas cuando los(as) oyentes proporcionan una respuesta que corresponde con la realidad y la línea roja, cuando no coincide con la realidad.

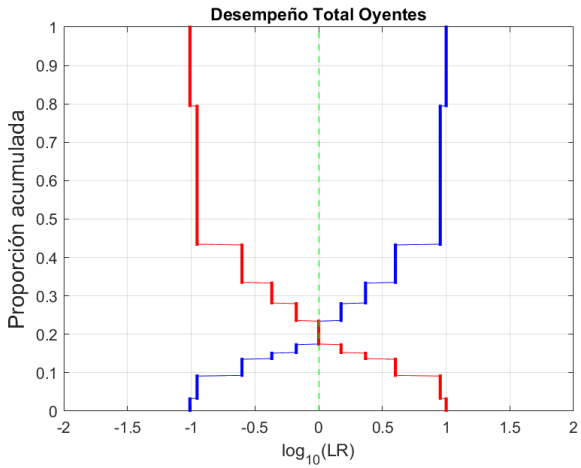

Figura 1: Gráfico de Tippet del desempeño general de todos(as) los(as) oyentes

Entre las muestras de audio de cada hablante, en algunos casos, hubo gran variabilidad. El caso más llamativo fue de Bachelet, quien tuvo el reconocimiento más débil entre los(as) hablantes, sin embargo, un audio real (R1, en la Matriz) que obtuvo un C_{llr} de 0.22, el más bajo de todos. Esto implicó que de todos(as) los(as) oyentes que respondieron nadie se equivocó en distinguirlo. La Figura 2 muestra un gráfico de Tippett con este resultado.

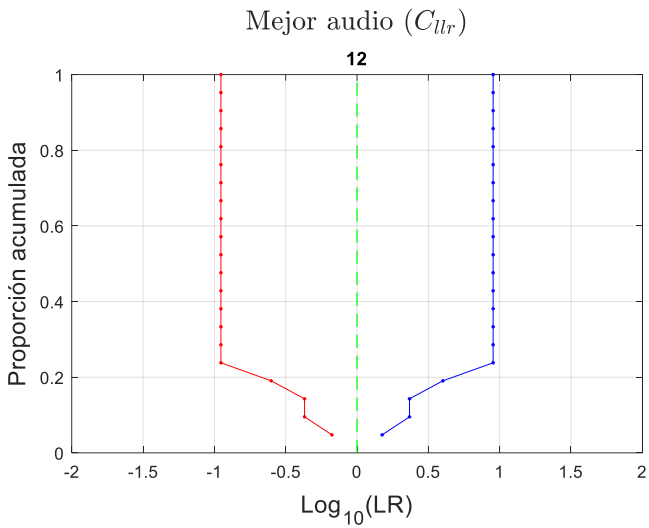

Figura 2: Gráfico de Tippett de un audio de voz real (R) de Bachelet con el más alto reconocimiento de todos(as) los(as) oyentes.

En un sentido contrario, Boric proporcionó un audio IA (IA1, en la Matriz) que obtuvo un C_{llr} 1.52, el más alto, lo que implicó que tuvo la mayor cantidad de respuestas incorrectas (tuvo un reconocimiento cercano al 50%). La Figura 3 muestra un gráfico de Tippett con este resultado.

En general, Boric es el que peor desempeño tiene en audios IA. Entre real e IA es Bachelet. No tenemos una explicación para este resultado, pero notamos factores acústicos típicos de las grabaciones (como la reverberación, ruido de fondo, ambiente en general) podrían haber afectado a la forma en que los individuos percibían los audios, y en consecuencia a sus respuestas, vinculando conscientemente estas características a audios más reales.

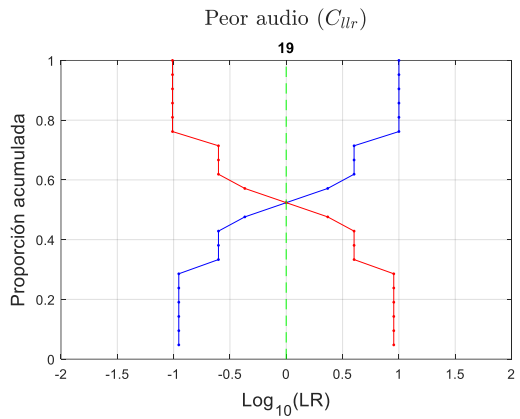

Figura 3: Gráfico de Tippett de un audio IA de Boric con el más bajo reconocimiento de todos(as) los(as) oyentes.

Entre los(as) oyentes también hubo variabilidad. El(la) oyente 39 obtuvo el C_{llr} más bajo (mejor desempeño) de 0.24. En tanto el (la) oyente 12 obtuvo el $C_{llr\,más}$ alto (peor desempeño) de 1.51. Las Figuras 4 y 5 muestran estos resultados.

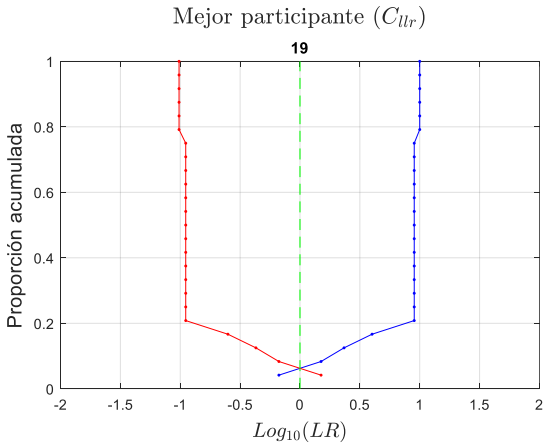

Figura 4: Gráfico de Tippett con el C_{llr} más bajo de 0.24 que corresponde al(la) oyente 39.

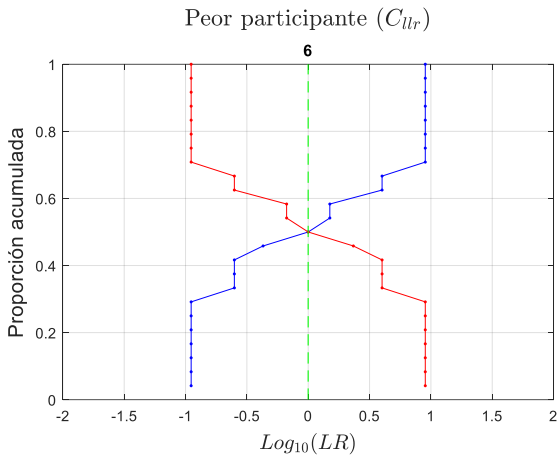

Figura 5: Gráfico de Tippett con el C_{llr} más alto de 1.51 que corresponde al(la) oyente 12.

La variabilidad entre oyentes es algo esperable, sin embargo, podríamos sospechar que el(la) oyente 39 hubiese tenido alguna experticia no declarada en la materia, criterio de exclusión, y a la inversa que el(la) oyente 12 no tuviera ningún conocimiento previo sobre voces clonadas, además de otros requisitos como familiaridad con las voces de interés (criterios de inclusión).

4. Conclusión

El desarrollo de la presente investigación ha permitido responder la pregunta central planteada:

> ¿Qué tan bien oyentes no expertos(as) distinguen voces familiares generadas por IA versus voces familiares generadas por humanos en el contexto potencial de una llamada fraudulenta?

Para responder esta pregunta usamos un enfoque tradicional que considera el conteo de aciertos y desaciertos, independientemente del grado en que las respuestas ser acercan o alejan del valor verdadero, y un enfoque moderno que sí recoge estas diferencias en términos de razones de verosimilitud (LR). Se evidenció que ambos enfoques se pueden complementar y contribuir en forma combinada a una mejor comprensión del desempeño auditivo.

La respuesta que obtuvimos a la pregunta planteada fue que los(as) oyentes no expertos(as) distinguen correctamente muestras de voces reales y clonadas a tasas muy altas.

El que oyentes no expertos(as) tengan un buen desempeño reconociendo voces familiares puede ser intuitivamente esperable; por el contrario, lo que se espera

genuinamente para las voces clonadas es un mal desempeño, y esto contrasta con los hallazgos de esta investigación.

Explicaciones plausibles de estos resultados incluirían factores relacionados con el número de muestras consideradas por cada hablante, el grado de familiaridad de los(as) oyentes con respecto a las voces presentadas y/o algún tipo de rasgo distintivo de las voces de interés que el software disponible no percibe, entre otros. Por otra parte, el hecho de que algunas muestras de voz hubieran sido más correctamente distinguibles que otras, independientemente del(la) hablante, podría estar relacionado específicamente con las características técnicas ambientales de grabación que dotarían de más o menos naturalidad a la muestra de interés considerada.

Todas las hipótesis mencionadas, que intentan proponer vías posibles de explicación de los hallazgos informados, deben ser probadas, por lo que se abren nuevas aristas de investigación.

OBRAS CITADAS

Arik, S., Chen, J., Peng, K., Ping, W., Zhou, Y. (2018). Neural Voice Cloning with a Few Samples. En S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol.31). CurranAssociates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/4559912e7a94a9c32b09d894f2bc3c82-Paper.pdf

Bali, A.S., Basu, N., Weber, P., Rosas-Aguilar, C., Edmond, G., Martire, K.A., Morrison, G.S. (2024). Speaker identification in courtroom contexts – Part III: Groups of collaborating listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. Forensic Science International, 360, 112048.https://doi.org/10.1016/j.forsciint.2024.112048

Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G., Khoury, E. (2020). Generalization of Audio Deepfake Detection, in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 132–137.

Edmond, G., San Roque, M. (2009). Quasi-justice: ad hoc expertise and identification evidence. Criminal Law J., 33, pp. 8-33.

Edmond, G., Martire, K., San Roque, M. (2011). Unsound law: issues with (expert') voice comparison evidence. Melbourne Univ. Law Rev., 35, pp. 52-112.

Geradts, Z., Franke, K. (Eds.), Artificial Intelligence (AI) in Forensic Sciences (pp. 3-20). Wiley.

Hughes, V., Rhodes, R. (2018). Questions, propositions and assessing different levels of evidence: Forensic voice comparison in practice. Science & Justice, 58(4), 250-257. https://doi.org/10.1016/J.SCIJUS.2018.03.007

IIElevenLabs (Aplicación de página web). [Software de audio]. [Fecha de último acceso:12/01/25]. https://elevenlabs.io/about

Laub, C.E., Wylie, L.E., Bornstein, B.H. (2013). Can the courts tell an ear from an eye? Legal approaches to voice identification evidence. Law Psychol. Rev. 37, pp. 119-158.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521, pp. 436-444. [Review]

Morrison, G:S., Enzinger, E., Zhang, C. (2018). Forensic speech science, in: I. Freckelton, H. Selby (Eds.), Expert Evidence (Ch. 99), Thomson Reuters, Sydney, Australia.

Moustafa, N. (2022). Digital Forensics in the Era of Artificial Intelligence (1st ed.). CRC Press. https://doi.org/10.1201/9781003278962

Morrison G.S., Enzinger E., Hughes V., Jessen M., Meuwly D., Neumann C., Planting S., Thompson W.C., van der Vloed D., Ypma R.J.F., Zhang C., Anonymous A., Anonymous B. (2021). Consensus on validation of forensic voice comparison. Science & Justice, 61, 229–309.https://doi.org/10.1016/j.scijus.2021.02.002

Mcuba, M., Singh, A., Adeyemi Ikuesan, R., Venter, H. (2023). The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation, Procedia Computer Science, Volume 219, pp. 211-219, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2023.01.283.(https://www.sciencedirect.com/science/article/pii/S1877050923002910)

Napolitano, D. (2020). The Cultural Origins of Voice Cloning. xCoAx. https://www.researchgate.net/publication/342924151

Ormerod, O. (2001). Sounds familiar? Voice identification evidence. Crim. Law Rev. (10), pp. 595–622.

Robson, J. (2018). 'Lend me your ears': an analysis of how voice identification evidence. is treated in four neighbouring criminal justice systems. Int. J. Evid. Proof 22, pp. 218–238. https://doi.org/10.1177/1365712718782989.

Rosas, C., Sommerhoff, J., Morrison, G.S. (2019). A method for calculating the strength. of evidence associated with an earwitness's claimed recognition of a familiar speaker. Science & Justice, 59, pp. 585-596

Rosas, C., Sommerhoff, J., Pacheco, J., Sáez, C. (2020). Yo lo reconocería por su voz… El caso de Emilio Berkhoff. Alpha (Osorno), (51), 137-160. https://dx.doi.org/10.32735/s0718-2201202000051851

Rose, P. (2002). Forensic Speaker Identification, Taylor and Francis, London UK.

Saini, K., Sonone, S.S., Sankhla, M.S., Kumar, N. (Eds.). (2024). Artificial Intelligence in Forensic Science: An Emerging Technology in Criminal Investigation Systems (1st ed.). CRC Press. https://doi.org/10.4324/9781003287810

Saleema, A., Thampi, S. M. (2018). Voice Biometrics: The Promising Future of Authentication in the Internet of Things. In Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science, IGI Global, pp. 360–389.

Simonite, T. (2022). A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be. Wired.

Solan, L.M., Tiersma, P.M. (2003). Hearing voices: speaker identification in court. Hastings Law J. 54, pp. 373–435.

Sherrin, C. (2016). Earwitness evidence: the reliability of voice identifications. Osgoode Hall Law J. 52, pp. 819–862. https://digitalcommons.osgoode.yorku.ca/ohlj/vol52/iss3/3.

Suguna, S.K., Dhivya, M., Paiva, S. (Eds.). (2021). Artificial Intelligence (AI): Recent Trends and Applications (1st ed.). CRC Press. https://doi.org/10.1201/9781003005629

Yarmey A.D., Yarmey A.L., Yarmey M.J., Parliament L. (2001). Commonsense beliefs and the identification of familiar voices. Applied Cognitive Psychology, 15, pp. 283–299. http://dx.doi.org/10.1002/acp.702

Yarmey, A.D. (2007). The psychology of speaker identification and earwitness memory, in: R.C.L. Lindsay, D.F. Ross, J.D. Read, M.P. Toglia (Eds.), The Handbook of Eyewitness Psychology, Memory for People, vol. II, Lawrence Erlbaum, Mahwah NJ, pp. 101-136. https://doi.org/10.4324/9781315805535.ch5.

Ypma R.J.F., Ramos D., Meuwly D. (2023). AI-based forensic evaluation in court: The desirability of explanation and the necessity of validation. In Geradts Z., Franke K. (Eds.), Artificial Intelligence (AI) in Forensic Sciences (pp. 3-20). Wiley.

Zawali, B., Ikuesan, R.A., Kebande, V. R., Furnell, S., A-Dhaqm, A. (2021). Realising a Push Button Modality for Video-Based Forensics. Infrastructures, vol. 6, no. 4, p. 54.

Zhao, L., Chen, F. (2020). Research on voice cloning with a few samples. *Proceedings - 2020 International Conference on Computer Network*, Electronic and Automation, ICCNEA 2020, 323-328. https://doi.org/10.1109/ICCNEA50255.2020.00073