XIII Congresso Ibérico de Acústica
55º Congreso Español de Acústica
TECNIACÚSTICA 2024
Faro, 11 a 13 setembro 2024

RECONOCIMIENTO PERCEPTIVO DE HABLANTES: UN EXPERIMENTO CON VOCES CLONADAS ARTIFICIALMENTE Y CON VOCES DE GEMELOS IDÉNTICOS

Eugenia San Segundo[1], Mark Gibson[2]

[1] Phonetics Laboratory, CSIC
{eugenia.sansegundo@csic.es}
[2] Universidad de Navarra, Speech Laboratory
{mgibson@unav.es}

Resumen

En este estudio diseñamos un experimento de percepción de elección forzada múltiple (igual/diferente) con Praat utilizando 12 muestras de voz diferentes (3-5 segundos cada una), extraídas de conversaciones espontáneas semidirigidas entre parejas de gemelos. Hay dos muestras de voz diferentes por cada pareja de gemelos, y cada voz también fue clonada artificialmente. El experimento consta de 20 estímulos en 3 tipos de emparejamientos: 8 del mismo hablante, 8 de hablantes diferentes y 4 combinaciones de un hablante con su voz *deepfake*. En total, 30 oyentes participaron en el experimento de percepción en condiciones controladas (mismos auriculares, ordenador y sala silenciosa). Los oyentes eran estudiantes universitarios, hablantes nativos de español peninsular estándar. No sabían que los estímulos incluían voces de gemelos y *deepfakes*. El objetivo de esta investigación fue determinar si los resultados de identificación, así como los tiempos de reacción, dependen del tipo de combinación de estímulos: mismo hablante, diferente hablante o real-*deepfake*.

Palabras clave: *deepfakes*, voz clonada, gemelos, prueba perceptiva

Abstract

We designed a multiple forced choice (same / different) listening experiment with Praat using 12 different voice samples (3-5 seconds long), extracted from semi-directed spontaneous conversations of two pairs of twins. There are two different voice samples per twin member, and each voice was also artificially cloned. The experiment was set up in Praat with 20 stimuli in 3 types of pairings: 8 same-speaker pairings, 8 different-speaker pairings, and 4 combinations of a speaker with his voice deepfake. 30 listeners took part in the perceptual experiment under controlled conditions (same headphones, computer, and silent room). Listeners were college students, native speakers of Standard Peninsular Spanish. They were not informed that the stimuli included twin voices and deepfakes. The objective of this investigation was to determine whether identification results, as well as reaction times, depend on the type of stimuli combination: same-speaker, different-speaker, or real-deepfake combinations.

Keywords: deepfakes, voice cloning, twins, perception test.

PACS nº. 43.71.-k, 43.71.Bp, 43.72.Uv

1 Introducción

A lo largo de los últimos años, diferentes medios han mostrado innumerables ejemplos de *deepfakes* de voz. Nos referimos a voces sintéticas generadas mediante modelos de aprendizaje profundo, que presentan un parecido extremo con las voces reales. De forma más amplia, hablamos de ataques de suplantación de identidad (*spoofing attacks* o *presentation attacks*) para referirnos a ataques destinados específicamente a engañar a un sistema biométrico, por ejemplo, de acceso por voz. La verificación automática del hablante (VAH) tiene como objetivo verificar la identidad de una persona usando su voz. Al igual que otros sistemas biométricos, la VAH es vulnerable a la suplantación de identidad a través de diversos vectores de ataque, incluida la clonación de voz.

La creciente popularidad de las redes neuronales profundas ha hecho que estos *deepfakes* estén cada vez más extendidos. Si bien estos avances tecnológicos pueden aportar enormes beneficios en el ámbito clínico (clonación de voz para pacientes con enfermedades neurodegenativas, por ejemplo) y en diversas aplicaciones tecnológicas, como dotar de características naturales a los asistentes de voz, si caen en las manos equivocadas, estos desarrollos ponen en serio peligro el uso de la voz con fines biométricos en el ámbito de la seguridad informática, así como el uso de muestras de voz como prueba forense en el ámbito legal y judicial (p.ej. a la hora de realizar cotejos de voces) [1-3] . A estas dos amenazas se suma una tercera, de gran relevancia social: el auge de las ciberestafas mediante el procedimiento conocido como *vishing* o fraude por voz; esto es, a través de una llamada telefónica, se suplanta la identidad de una empresa, organización o persona de confianza, con el fin de obtener información personal y sensible de la víctima, incluyendo información financiera. Los casos más paradigmáticos son aquellos en los que la voz de la llamada telefónica es de un parecido extremo con algún familiar de la víctima de la estafa, a la que se urge a que haga un envío de dinero.

Igualmente, los *deepfakes* se utilizan para difamar a figuras públicas (p. ej. políticos) y hacerles emitir mensajes falsos, con el fin de influir en elecciones o en decisiones políticas. Así, cada vez se hace más difícil distinguir las noticias reales de las falsas y, por ello, estamos siendo testigos de una falta de confianza sin precedentes en los medios. Por lo tanto, es de suma importancia implementar una metodología que identifique cómo distinguir qué muestras de voz son *deepfakes* y cuáles son reales. Esto permitirá proteger a los ciudadanos y construir sociedades más justas y seguras (Objetivo 16 de la Agenda 2030 para el Desarrollo Sostenible de Naciones Unidas).

Algunos estudios recientes examinan el estado de la cuestión en detección de ataques de suplantación de voz. Por ejemplo, en [4] se revisaron 172 artículos publicados entre 2015 y 2021 con el objetivo de analizar sistemáticamente el estado actual de las investigaciones en detección de ataques de suplantación de voz, proporcionando una taxonomía útil de los tipos de ataques que podemos identificar y los problemas que todos ellos comparten, además de resaltar direcciones futuras de trabajo en este ámbito. Por su parte, [5], en un estudio más reciente, aportan una revisión exhaustiva, en la que se señalan las diferencias clave entre los distintos tipos de *deepfakes* de voz, se describen y analizan conjuntos de datos disponibles, características acústicas analizadas, así como tipos de clasificaciones y métricas de evaluación, junto con una descripción de los enfoques metodológicos de última generación. Para cada aspecto se discuten las técnicas básicas, los últimos desarrollos y los principales desafíos. Además, estos autores realizan una comparación unificada de características y clasificadores representativos en los distintos conjuntos de datos para la detección de *deepfakes* de audio. Su estudio muestra que las investigaciones futuras deberían abordar la falta de conjuntos de datos a gran escala, la mala generalización de los métodos de detección existentes para ataques falsos de naturaleza desconocida, así como la interpretabilidad de los resultados de la detección.

Con todo, escasean los estudios que investiguen las capacidades humanas para detectar *deepfakes*. Resumimos a continuación los principales resultados de las investigaciones más recientes que hemos encontrado. El estudio de [6] parte de la idea de que conocer cómo identifican *deepfakes* los humanos puede conducir a una mejor comprensión del funcionamiento de los sistemas de aprendizaje automático que detectan *deepfakes* y que funcionan como cajas negras, por estar basados en redes neuronales profundas. Los autores de este reciente estudio ([6]) diseñan dos condiciones experimentales. Por un lado, presentan a un grupo de oyentes un audio y les preguntan si consideran que el audio es falso (esto es, un *deepfake*). Por otro lado, presentan a otro grupo de oyentes pares de audios que contienen la misma voz (una es la voz natural, también llamada *bona fide*, y otra sintética) y les piden que identifiquen la voz sintética. Además, este estudio se llevó a cabo con grabaciones en dos lenguas: inglés y mandarín, con el fin de observar si los oyentes utilizan características específicas de la lengua para detectar *deepfakes* y para investigar si la tarea es más fácil en una lengua que en otra. Por último, este estudio incorpora en algunos test una fase de familiarización con ejemplos de voces sintéticas para comprobar hasta qué punto este tipo de intervención aumenta el porcentaje de aciertos. En términos generales, los resultados de este estudio muestran que los humanos podemos detectar *deepfakes* perceptivamente, pero no de forma consistente o fiable. Los participantes de este estudio identificaron correctamente los *deepfakes* un 73 % del tiempo. Otros resultados y conclusiones interesantes de la investigación descrita en [6] son:

- No existe diferencia entre la capacidad de detección de *deepfakes* de voz en inglés y la capacidad de detección de *deepfakes* cuando estos se presentan en mandarín.
- Familiarizar a los oyentes con ejemplos de *deepfakes* mejora su capacidad de detección, pero solo ligeramente.
- En los audios más cortos no es más fácil identificar *deepfakes*.
- Escuchar los estímulos más veces no favorece la detección de *deepfakes*.
- Pasar más tiempo realizando el experimento perceptivo no afecta a los resultados.
- Los participantes que clasificaron correctamente los estímulos *bona fine* mencionan que se fijaron en las pausas, el tono y la entonación. Sin embargo, los participantes que categorizaron incorrectamente los estímulos *bona fide* como falsos también se refirieron a esos mismos atributos como pistas que les ayudaron a decidir.
- Los participantes tienden a confiar en su intuición para tomar sus decisiones perceptivas, refiriéndose a la naturalidad y el timbre robótico de los audios. Más allá de la intuición, los participantes en inglés y mandarín también hacen referencia a las pausas, la entonación, la pronunciación y la velocidad.
- En cuanto a las diferencias entre idiomas, hubo más referencias a la respiración por parte de los participantes de habla inglesa. Por el contrario, los participantes que hablaban mandarín mencionaron la cadencia del hablante, el espacio entre palabras y la fluidez. Estas diferencias, según los autores [6], podrían deberse a las diferencias rítmicas entre las dos lenguas.

Además de los resultados del reciente estudio llevado a cabo por [6], que acabamos de resumir, nuestra revisión bibliográfica sobre percepción de *deepfakes* revela lo siguiente:

- Existen muy pocos estudios que examinen la capacidad humana de detectar *deepfakes* [7,8,9]
- De los estudios existentes, la mayoría se centran en *deepfakes* de imágenes o de vídeo (estímulos multimodales), pero no de voz [10, 11].
- Algunos estudios previos apuntan a que en los audios más cortos es más fácil detectar los *deepfakes* de voz [7], a diferencia de los resultados de [6].
- Otros estudios han encontrado que la diferencia entre la precisión humana y la de un sistema de inteligencia artificial para detectar *deepfakes* es del 10 % [9].

2 Método

2.1 Hablantes

Se eligió a dos parejas de gemelos varones idénticos (es decir, monocigóticos) del Corpus de Gemelos descrito en [12] (hablantes monolingües de la variedad de español peninsular estándar). Se establecieron dos criterios principales para seleccionar a las dos parejas de gemelos del corpus: (i) edad similar (20 y 28 años cada pareja) y (ii) distancia euclidiana (DE) similar entre cada hablante y su gemelo. Las DE se habían calculado en un estudio previo [13] y están basadas en la evaluación perceptiva de su cualidad de voz mediante una versión simplificada del Análisis del Perfil Vocal (*Vocal Profile Analysis, VPA*) [14]. Las dos parejas de gemelos elegidas presentan, por tanto, un *Similarity Matching Coefficient* de 0.8, de un máximo posible de 1. Esto quiere decir que se asemejan perceptivamente en 8 de 10 ajustes articulatorios posibles del VPA simplificado descrito en [15].

2.2 Estímulos

El test perceptivo consta de 20 estímulos. Para ello se utilizaron 12 muestras de voz diferentes (3-5 segundos cada una), extraídas de conversaciones espontáneas semidirigidas entre parejas de gemelos, que fueron seleccionados con los criterios que explicamos en el apartado 2.1. Hay dos muestras de voz diferentes por cada pareja de gemelos, y cada voz también fue clonada artificialmente. Los 20 estímulos que conforman el test se obtienen de 3 tipos de emparejamientos de voces: 8 del mismo hablante, 8 de hablantes diferentes y 4 combinaciones de un hablante con su voz *deepfake*.

En lo que respecta al contenido de las grabaciones y al estilo de habla, se trata de fragmentos de conversaciones espontáneas semidirigidas que entabló individualmente cada gemelo con el Autor 1 [16]. El interlocutor es una variable controlada, lo que resulta en el mismo tipo de estilo de habla en todas las conversaciones. Todas las frases son enunciados declarativos de diversos temas neutros.

2.3 Oyentes

En el test perceptivo participaron 30 hablantes nativos de español (rango de edad 18-39). Todos ellos son estudiantes o empleados de la Universidad de Navarra. Ninguno presentaba dificultades auditivas**.**

2.4 Diseño del experimento perceptivo

Se diseñó en Praat [17] un experimento de elección forzada múltiple con 20 emparejamientos de estímulos, que se presentaron a los oyentes en orden aleatorio. Los oyentes debían indicar cada vez si las voces eran las mismas o diferentes. A los participantes no se les informó en ningún momento de que los estímulos incluían voces de gemelos o *deepfakes*. La prueba se realizó en un PC ubicado en una habitación silenciosa. Los oyentes realizaron la prueba con auriculares. El tiempo de reacción (tiempo que tarda cada oyente en tomar una decisión) se midió desde el final del segundo estímulo. La duración de la prueba fue aproximadamente de 10 minutos.

2.5 Análisis estadístico (test de percepción)

Para el test de percepción consideramos dos variables de respuesta: *precisión* (que mide aciertos en las respuestas; dos niveles: correcto e incorrecto) y el tiempo de reacción (el tiempo que va desde). *Condición* (tres niveles: M=mismo hablante, D=distinto hablante, MF=*deepfake*) y *estímulo* (11

combinaciones de hablantes * dos gemelos = 22) eran los predictores principales, aunque también modelamos sexo biológico del participante y otros datos personales que pudieran tener un efecto en las respuestas de los participantes, como el origen del hablante. Para modelar la *precisión* utilizamos modelos mixtos lineales generalizados (en adelante, GLMM) con el paquete "lme4" [18] en RStudio [19]. Se empleó esta clase de modelo estadístico por el carácter binario de la variable de respuesta (correcto o incorrecto). Codificamos "participante" como efecto aleatorio, ya que esperamos variación entre los participantes que no está relacionada con el predictor. La máxima verosimilitud $\chi 2$ está basada en la desviación estadística para determinar significación.

Ya que el tiempo de reacción es una variable continua, utilizamos modelos lineales de efectos mixtos con el paquete "lme4" en RStudio. De nuevo, el predictor principal de este modelo es *condición* (tres niveles: M=mismo hablante, D=distinto hablante, MF=*deepfake*).

3 Resultados

En cuanto a la variable de respuesta *precisión,* que compara la cantidad de aciertos (han respondido 'misma persona' cuando los dos audios eran de la misma persona, y han respondido 'distinta persona' cuando los audios eran de distintas personas), solo ha habido un efecto de *condición* para la condición M (respuesta de 'misma persona') (N=660, p<0.001, véase la Tabla 1). Para D (respuesta de 'distinta persona') y DF (respuesta de 'misma' persona cuando una de las voces era de un *deepfake*) los resultados son parecidos, lo cual indica la ausencia de sistematicidad en las respuestas en función de la condición/estímulo. Estos resultados muestran que los participantes eran capaces de identificar mejor a los mismos hablantes (acierto superior para identificar en M cuando eran las mismas personas), pero entre los *mismatch* con el gemelo real y el *deepfake*, no ha habido ningún efecto (véase la Tabla 1). En cuanto a los estímulos individuales, no ha habido ningún efecto de la variable *precisión* con un estímulo (o un tipo de estímulo) en concreto (véase la Tabla 2).

Tabla 1 - Resumen de los parámetros de regresión estimada del predictor *condición*: *Estimate, standard error (SE), z-ratio y valor p*

Condición	Estimate	*SE*	*z-ratio*	*p*
Intercept (D)	-0.10	0.32	-0.32	0.75
M	-2.10	0.61	-3.41	<0.001
MF	-0.52	0.57	-0.92	0.36

Nota. Niveles de significancia. * p < .05, ** p < .01, *** p < .001

Tabla 2 - Resumen de los parámetros de regresión estimada del predictor *estímulo*: *Estimate, standard error (SE), z-ratio y valor p.* Los estímulos que contienen un *deepfake* están sombreados en gris.

Estímulo	Estimate	Std. Error	z value	p
stimulus01Aa,01Af	0.98	1.44	0.68	0.49
stimulus01Aa,01Ba	0	1.29	0	1
stimulus01Ab,01Aa	-37.19	30	0	1
stimulus01Ab,01Bb	54.61	30	0	1
stimulus01Ba,01Aa	0	1.29	0	1
stimulus01Ba,01Bb	-30.19	27	0	1
stimulus01Ba,01Bf	0.98	1.44	0.68	0.49
stimulus01Bb,01Bf	-9.80	0	-0.680	0.497
stimulus01Bb,01Ab	1.79	1.44	1.24	0.21
stimulus01Bb,01Ba	-0.29	1.60	-0.18	0.86
stimulus02Aa,02Ab	-33.05	11	0	1
stimulus02Aa,02Ba	-0.98	1.44	-0.68	0.49
stimulus02Ab,02Aa	-28.44	10	0	1
stimulus02Ab,02Af	0.98	1.44	0.68	0.49
stimulus02Ab,02Bb	0	1.29	0	1
stimulus02Ba,02Aa	-0.98	1.4	-0.68	0.49
stimulus02Ba,02Bb	0.69	1.5	0.47	0.64
stimulus02Bb,02Ba	-30.76	37	0	1

Nota. Niveles de significancia. * $p < .05$, ** $p < .01$, *** $p < .001$

La Tabla 1 resume los parámetros de regresión estimada del predictor *condición*, mientras que los resultados para *estímulo* aparecen en la Tabla 2: *Estimate, standard error (SE), z-ratio y valor p.*

A nivel general, los participantes no mostraron mucha dificultad en el test de percepción. Un 70 % de las respuestas (N=462) son correctas. De este 70 %, hubo más aciertos cuando las voces pertenecían al mismo hablante: 51.4 % del 70 % total (N=237). En cuanto a la capacidad de discernir entre voces distintas, el 30 % (N=138) contestó acertadamente 'distinto hablante' cuando las voces eran de dos hablantes distintos. Los que contestaron correctamente 'distinto hablante' cuando los estímulos eran *deepfakes* constituyen el restante 19 % de los aciertos.

En cuanto a las respuestas incorrectas, es decir, cuando los participantes respondieron 'mismo hablante' cuando los estímulos eran de dos hablantes distintos o viceversa, el 63.3 % (N=125) eran para la condición 'distinto hablante', seguido por la condición *deepfake* (17.5 %), y por 'mismo hablante' (13 %).

El conjunto de estos resultados indica que, por lo general, los participantes podían percibir cuándo los estímulos eran del mismo hablante y cuándo eran de hablantes distintos. Presentaron más dificultad a la hora de discernir si eran distintos hablantes que mismo hablante. No existe mucha sistematicidad en cuanto a la detección de los *deepfakes*. Hubo más respuestas correctas (es decir, respondieron más veces que eran hablantes distintos que hablantes diferentes), pero la diferencia no es significativa estadísticamente.

En cuanto a los tiempos de reacción, los resultados de nuestros modelos lineales de efectos mixtos muestran un efecto principal para el predictor *condición* ($\chi 2$ [8.41, N = 660] = 18.42, p = 0.015), pero no ha habido ningún efecto en cuanto al *estímulo* ($\chi 2$ [.001, N = 660] = 0.001, p = 0.99).

Los participantes presentaron tiempos de reacción parecidos cuando los estímulos eran del mismo hablante (M=8.28 s., sd=1.66 s.) y de hablantes distintos (M=8.72 s., sd=1.52 s.). Sin embargo, cuando los participantes escucharon estímulos en los que una de las voces era un *deepfake*, el tiempo de reacción medio aumentó de manera significativa (M= 9.8 s., sd= 3 s.).

4 Discusión y conclusiones

El hecho de que los participantes presenten tiempos de reacción parecidos cuando los estímulos son voces naturales, tanto del mismo hablante como de hablantes distintos, mientras que el tiempo de reacción medio aumenta significativamente cuando uno de los estímulos es un *deepfake*, podría indicar que existe un esfuerzo cognitivo mayor cuando el oyente se enfrenta a una voz sintética.

Por otro lado, es ampliamente conocido que el procesamiento de voces familiares y de voces no familiares es diferente. En el modelo de reconocimiento de voces familiares propuesto por Lavner et al. [20], las voces familiares se representan en términos de su desviación acústica de un prototipo de voz. Este prototipo se describe como un promedio de todas las voces que ha encontrado un oyente en su vida. Cada vez que escucha una voz, se calculan las características que se desvían del prototipo y el resultado de esta extracción de características se compara con patrones de referencia existentes o representaciones de voces conocidas. Si la distancia entre las características que se desvían de la voz percibida y el patrón de referencia es suficientemente pequeña, se reconocerá la voz como perteneciente a una persona familiar concreta. En este modelo no se especifica el funcionamiento de la percepción de voces no familiares. Aunque el modelo incluye un bucle para indicar que las voces percibidas que no se ajustan a un patrón de referencia inicial se comparan iterativamente con otros patrones de referencia,

(presumiblemente hasta que se encuentra la voz coincidente y se reconoce la voz familiar), no se propone realmente ningún mecanismo para explicar cómo se procesan como desconocidas las voces verdaderamente desconocidas. En un estudio reciente, Lavan y McGettigan [21] repasan otros modelos existentes [22, 23] y presentan una propuesta más amplia, en la que los oyentes tienen un objetivo perceptivo común de percibir a quién están escuchando, tanto si la voz les resulta familiar o desconocida. Se trata del modelo de Percepción de Personas a partir de Voces (PPV), que establece que los oyentes logran este objetivo a través de un mecanismo común de reconocimiento de una persona familiar, un personaje o un conjunto de características del hablante. Su modelo tiene como objetivo proporcionar una explicación más completa de cómo los oyentes perciben a la persona que están escuchando, utilizando un enfoque que incorpora aspectos de los marcos jerárquicos y mecanismos basados en prototipos propuestos dentro de los modelos anteriores de reconocimiento de voces.

Generalmente, los estudios que se centran en voces familiares y no familiares no han incluido *deepfakes* en sus trabajos. Por voces familiares, autores como los citados anteriormente [20 - 23], se refieren a voces de amigos, familiares, pareja, etc. Una voz no familiar sería aquella que escuchamos por primera vez; esto es, la voz de un extraño. En el experimento que hemos diseñado para esta investigación, obviamente todas las voces (las naturales y las artificiales) entrarían dentro de la etiqueta de voces "no familiares" para nuestros oyentes. Sin embargo, se sabe poco del procesamiento neuronal de las voces sintéticas. Los resultados obtenidos para la variable tiempo de reacción en esta investigación apuntan, por un lado, a la necesidad de realizar más investigaciones desde una perspectiva neurológica, con el fin de explorar de qué manera procesamos las voces falsas. Esto incluye enfoques científicos que hasta ahora no se habían explorado. Por ejemplo, en el ámbito de los *deepfakes* de imágenes, Li et al. [24] midieron los movimientos oculares de varios sujetos para determinar qué miraban y durante cuánto tiempo cuando se exponían a *deepfakes* y a imágenes reales, y encontraron que la forma en la que los ojos se mueven en un caso y en otro es diferente.

Tanto los resultados de [24] en el ámbito de los *deepfakes* de imagen como nuestros resultados sobre tiempo de reacción con *deepfakes* de voz parecen apuntar a un procesamiento distinto de los estímulos falsos frente a los naturales, lo que abre la puerta a que el ser humano sea capaz de detectar este tipo de material audiovisual manipulado. Pese a los avances tecnológicos y al alto grado de naturalidad que presentan los *deepfakes* hoy en día, todavía existe algo, por sutil que sea, que hace que reaccionemos de forma distinta cuando lo que percibimos con nuestros sentidos no es humano.

En estudios futuros, y en la línea de [6], analizaremos las respuestas cualitativas de los participantes de nuestro test perceptivo para conocer en qué aspectos de las voces se fijaron estos para tomar sus decisiones en este experimento. Por ejemplo, en el estudio que describimos ampliamente en la introducción ([6]), se mencionan pausas, tono y entonación. Sin embargo, su estudio se realizó con voces en inglés y en mandarín. Sería interesante averiguar cuáles son las claves acústicas que utilizan los oyentes en español, particularmente en un caso tan desafiante como este, que incluye los extremos de la similitud humana (voces de gemelos idénticos). Igualmente, analizaremos acústicamente los estímulos —naturales y sintéticos— utilizados en este experimento para evaluar comparativamente las características acústicas de unos y otros.

Fuentes de financiación

Proyecto PID2021-124995OA-I00 financiado por MICIU/AEI /10.13039/501100011033 y por FEDER, UE.

Referencias

[1] Brewster, T. (2021). Fraudsters Cloned Company Director's Voice in $35 Million Bank Heist, Police Find. *Forbes Magazine*. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-frauduses-deep-fake-voice-tech-to-steal-millions/?sh=75dd610b7559

[2] Chivers, T. (2019). What do we do about deepfake video? *The Guardian.* https://www.theguardian.com/technology/2019/jun/23/whatdo-we-do-about-deepfake-video-ai-facebook

[3] San Segundo, E. (2024). Profundizando en los *deepfakes*: ¿Qué hace humana a una voz? *Anuario AC/E 2024 de Cultura Digital* (pp. 28-41), Acción Cultural Española (AC/E).

[4] Tan, C.B., Hijazi, M.H.A., Khamis, N., Zainol, Z., Coenen, F., & Gani, A. (2021). A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction. *Multimedia Tools and Applications 80* (21), 32725–32762.

[5] Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023). Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970.*

[6] Mai, KT., Bray, S., Davies, T., & Griffin, LD. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLoS ONE 18*(8): e0285333.

[7] Watson, G., Khanjani, Z., & Janeja, VP. (2021). Audio Deepfake Perceptions in College Going Populations. *arXiv preprint arXiv:211203351*.

[8] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., et al. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, *64*:101114.

[9] Müller, NM., Pizzi, K., & Williams, J. (2022). Human perception of audio deepfakes. In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedi*a; pp. 85–91.

[10] Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences, 119*(1): e2110013119.

[11] Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., et al. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-16.

[12] San Segundo, E. (2013). A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application. *Procedia - Social Behav. Sci.*, 95, 59-67.

[13] San Segundo, E., Foulkes, P., Hughes, V. (2016). Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli. In *Proceedings of the 16th Australasian Conference on Speech Science and Technology* (pp. 309-312).

[14] Laver, J. (1980). The Phonetic Description of Voice Quality, Cambridge University Press.

[15] San Segundo, E., & Mompean, J. A. (2017). A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity. *Journal of Voice, 31(*5), 644-e11.

[16] San Segundo, E. (2014). *Forensic speaker comparison of Spanish twins and non-twin siblings. A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics.* Doctoral Dissertation, Consejo Superior de Investigaciones Científicas & Universidad Internacional Menéndez Pelayo.

[17] Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer* [Computer software] (Version 5.3.79).

[18] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Version 1.1.29. *Journal of Statistical Software*, 67, 1, 1−48. https://doi.org/10.18637/jss.v067.i01

[19] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
[20] Lavner, Y., Rosenhouse, J. & Gath, I. (2001). The prototype model in speaker identification by human listeners. *Int. J. Speech Technol. 4*, 63–74.
[21] Lavan, N., & McGettigan, C. (2023). A model for person perception from familiar and unfamiliar voices. *Commun Psychol 1*, 1.
[22] Belin, P., Bestelmeyer, P. E. G., Latinus, M. & Watson, R. (2011). Understanding voice perception. *Br. J. Psychol.* 102, 711–725.
[23] Belin, P., Fecteau, S. & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* 8, 129–135.
[24] M. Li, B. Liu, Y. Hu & Wang, Y. (2021). Exposing Deepfake Videos by Tracking Eye Movements, *25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 5184-5189.