PROFUNDIZANDO EN LOS *DEEPFAKES:* ¿QUÉ HACE HUMANA A UNA VOZ?

Eugenia San Segundo Fernández

Eugenia San Segundo Fernández es científica titular del CSIC, en el área de fonética experimental y aplicada. Licenciada en Filología Hispánica y Filología Inglesa (Universidad de Salamanca), con un doctorado en Estudios Fónicos (CSIC-UIMP), ha sido investigadora posdoctoral en la Universidad de York y profesora en la UNED. Actualmente dirige el laboratorio de fonética del Centro de Ciencias Humanas y Sociales del CSIC, lidera un proyecto de investigación sobre *deepfakes* y es la directora de la revista científica *Loquens: Spanish Journal of Speech Sciences* (CSIC).

1. La irrupción de la inteligencia artificial

La expresión *inteligencia artificial* (IA) ya forma parte de nuestro día a día. La escuchamos en el telediario y en la radio, pero también está en las redes sociales y en cualquier otra forma de comunicación, también las no digitales. Surge hasta en una conversación anodina con tu vecino, cuando este te comenta que los móviles nos escuchan, porque a él le han aparecido anuncios de cierta marca en su Instagram después de hablar con su madre sobre los regalos de Navidad para la familia: «Esto es cosa de la IA».

No es de extrañar que en 2022 la expresión compleja *inteligencia artificial* resultara elegida expresión del año por la Fundación del Español Urgente (FundéuRAE), promovida por la Agencia EFE y la Real Academia Española. La FundéuRAE, que otorga cada año el título de palabra del año a una palabra o expresión, así lo explica en una entrada de su página web,[1] con fecha 29 de diciembre de 2022:

> Esta construcción está definida en el diccionario académico como 'disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico'.

Hay que destacar que la expresión no es nueva. Según la propia FundéuRAE, se incorporó al diccionario de la Academia en su edición de 1992. Treinta años después, la FundéuRAE la selecciona por dos motivos: su destacada presencia en los medios de comunicación y el debate social provocado por los avances desarrollados en el ámbito de la IA y sus consecuencias éticas derivadas.

Algunos de estos debates éticos los resumía el programa *Equipo de Investigación* en su reportaje «Inteligencia artificial: Desafío total» del pasado 10 de noviembre. Por ejemplo, sabemos que ChatGPT nos permite traducir textos, sintetizar documentos y buscar información. No es de extrañar que millones de personas lo utilicen a diario, desde estudiantes a médicos, pasando por periodistas, profesores, investigadores, etc. El problema comienza cuando nos paramos a reflexionar en que muchos de esos trabajos hasta ahora los llevaban a cabo únicamente humanos; esto es, profesionales de distintos ámbitos que empiezan a ver amenazado su puesto de trabajo.

Está claro que esta aplicación de inteligencia artificial generativa ha revolucionado la forma de trabajar... por escrito. Pero ¿qué ocurre con la comunicación oral? En esta publicación nos centraremos en ese otro aspecto de la comunicación humana que también ha trastocado la inteligencia artificial: la voz.

Cuando el 15 de mayo de 2023 la asociación española La General de Locutores publica en su página web un comunicado de clausulado IA,[2] la sociedad empieza a tomar conciencia de la preocupación existente en el sector de los actores y actrices de doblaje y de voz —junto con el colectivo de locutores— por la implantación de la IA en su sector profesional.

En este comunicado, PASAVE (Plataforma de Asociaciones y Sindicatos de Artistas de Voz de España) exige un marco de seguridad jurídica que preserve los derechos de los artistas y profesionales de la voz, conscientes de que la cesión de los derechos sobre su voz e interpretación no estaban destinados originalmente a las finalidades derivadas de la nueva realidad que la IA entraña; esto es, entrenar sistemas de IA sin contraprestación alguna. En la figura 1 mostramos el contenido de las dos cláusulas que en este comunicado se propone que los profesionales del

[1] https://www.fundeu.es/recomendacion/inteligencia-artificial-es-la-expresion-del-2022-para-la-fundeurae/

[2] https://lageneraldelocutores.es/comunicado-de-clausulado-ia/

sector incluyan en todas las cesiones de derechos en doblaje a partir del 1 de enero de 2024.

> No se permite ni cede el uso de la voz, modulación, timbre, gestos y análogos del locutor/de la locutora y/o del actor/de la actriz de doblaje o de voz, para ser utilizados para alimentar, entrenar, simular o acciones similares, en programas o proyectos de inteligencia artificial (IA), robótica o cualquier metodología que utilice o transforme la voz grabada originalmente por el locutor/la locutora y/o el actor/la actriz, para destinarse a otro fin que el detallado en este contrato, que es dar voz e interpretar a un/os personaje/s de una producción en concreto. Solo podrá ser distorsionada, en su caso, para esa misma producción.
>
> No se permite ni cede el uso de la voz, modulación, timbre, gestos y análogos del locutor/de la locutora y/o del actor/de la actriz de doblaje o de voz, para ser utilizados para alimentar, entrenar, simular o acciones similares, en programas o proyectos de inteligencia artificial (IA), robótica, juegos informáticos o cualquier metodología que utilice o transforme la voz e interpretación grabada originalmente por el locutor/la locutora y/o el actor/la actriz, para destinarse a otro fin distinto al detallado en este contrato.

Figura 1. Cláusulas IA de PASAVE (Plataforma de Asociaciones y Sindicatos de Artistas de Voz de España).

La situación que desencadenó este comunicado fue el encargo de una multinacional a un estudio de grabación español para que organizara una «convocatoria de emociones» en la que los actores debían expresar con su voz distintas entonaciones y sentimientos. Como explica el presidente del Sindicato de Artistas de Doblaje de Madrid en esta entrevista,[3] pararon la convocatoria cuando fueron conscientes de que sus voces se iban a usar para entrenar una red neuronal.

En un mundo globalizado, es evidente que las preocupaciones de los artistas de la voz en España las comparten sus colegas en el resto de Europa, Estados Unidos, Latinoamérica, etc. Precisamente, United Voice Artists, UVA (Artistas de Voz Unidos), es una agrupación de asociaciones y sindicatos de profesionales de la voz en todo el mundo que nace con el fin de proteger y preservar el patrimonio artístico de locutores y actores de voz y de doblaje. Como se indica en su página web,[4] actualmente engloba a más de diez países con el fin de participar en la toma de decisiones para establecer normativas que regulen el uso de la IA para proteger los derechos de propiedad intelectual de los artistas de la voz y velar por el cumplimiento del Reglamento General de Protección de Datos (RGPD). Según recoge su propio manifiesto, hacen un llamamiento a los políticos y legisladores de la Unión Europea para que aborden, en otros aspectos, los riesgos inherentes, tanto legales como éticos, en la concepción, entrenamiento y comercialización del contenido generado por IA. Un punto fundamental de su manifiesto, que resume las preocupaciones de este sector, es el siguiente:

> En la actualidad, la tecnología de IA generativa depende en gran medida de las fuentes presentes online para mejorar sus capacidades de aprendizaje, lo que suele implicar el uso ilícito de datos y contenidos protegidos por derechos de autor. Este proceso de recopilación no se preocupa de verificar si estos datos y contenidos pueden reutilizarse o no.

Y una petición clara que se desprende de dicha preocupación, y que podemos leer en su manifiesto, reza así:

> Cualquier uso de la tecnología de IA para generar y clonar voces humanas debe estar sujeto al consentimiento explícito de los artistas e intérpretes de voz, que deben, por tanto, estar en condiciones de rechazar la utilización de sus obras e interpretaciones, pasadas y futuras, para fines no expresamente

3 https://www.publico.es/sociedad/inteligencia-artificial-entrena-robado-millones-artistas.html

4 https://unitedvoiceartists.com/

autorizados por ellos, y que se les ofrezcan soluciones prácticas para garantizar la eficacia de esta elección.

Un ejemplo de que la unión hace la fuerza es el caso de la reciente huelga de actores de Hollywood. Han sido necesarios 146 días de huelga —o, lo que es lo mismo, casi cinco meses— para llegar a un acuerdo entre la industria del cine y el sindicato de guionistas de Hollywood, que ha resultado en un contrato, firmado por ambas partes, con la totalidad de las demandas del sindicato satisfechas.

Los sistemas de IA generativa producen contenido sintético de audio, imagen, vídeo o texto que pueden aplicarse a muchas tareas diferentes.

Una de las peticiones tenía que ver precisamente con las restricciones al uso de la IA generativa. Como define la Comisión Europea,[5] los sistemas de IA generativa son sistemas de IA que generan contenido sintético de audio, imagen, vídeo o texto, y que pueden aplicarse a muchas tareas diferentes en diversos campos. Básicamente, lo que prohíbe el acuerdo al que han llegado los guionistas de Hollywood es que se utilice el trabajo de un guionista para entrenar una IA.

2. Del arte a la ciencia

2.1. Artistas y famosos afectados por la IA

Es evidente que existen usos de la IA que son inofensivos, además de útiles. Consideremos, por ejemplo, el último videoclip de Rosalía y Björk. El pasado 21 de noviembre se estrenaba la colaboración musical entre la artista española y la islandesa. En el videoclip se utilizó IA para hacer coincidir en tiempo y espacio a ambas cantantes, que protagonizan una especie de lucha con catanas. Puesto que es una canción que Björk ya había cantado hace veinticinco años (tenía casi la misma edad que tiene hoy Rosalía), el resultado que se consigue con este experimento realizado con IA es que en el videoclip la edad vocal de ambas artistas sea la misma. Según la fotógrafa y directora de arte que está detrás del vídeo,[6] en algunas tomas los rostros de las artistas están generados con IA. Según otras fuentes,[7] las imágenes de sus caras estarían directamente superpuestas a los rostros de dos expertas gimnastas, que son las que aparecen en ese combate virtual que vemos en el vídeo.

Sea como fuere, el videoclip de Björk y Rosalía es un caso de uso consentido de la IA con fines artísticos. Otro ejemplo habitual de uso consentido de la IA es aquel que se hace con fines clínicos para la clonación de voz en pacientes con trastornos neurodegenerativos. Por ejemplo, la esclerosis lateral amiotrófica (ELA) es una enfermedad degenerativa de tipo neuromuscular que conduce a la pérdida de la capacidad de hablar, tragar o respirar. Todos recordamos al astrofísico y divulgador Stephen Hawking, que fue diagnosticado de ELA a los veintiún años. Hasta hace poco, las voces sintéticas sonaban poco naturales, pues carecían de «personalidad». Sin embargo, gracias al empleo de la IA, hoy en día estas voces suenan considerablemente más auténticas, pues recrean la voz de una persona concreta mediante la clonación vocal.[8]

Con el fin de conservar la voz de estas personas, existen varias iniciativas científicas que utilizan herramientas de IA para almacenar las voces de las personas mientras aún pueden hablar y después «recrean» esas voces mediante tecnología

5 https://ec.europa.eu/commission/presscorner/detail/es/IP_24_85

6 https://www.revistavanityfair.es/articulos/carlota-guerrero-la-innovadora-creadora-detras-del-video-de-rosalia-y-bjork

7 https://los40.com/2023/11/21/rosalia-y-bjork-lanzan-oral-con-division-de-opiniones-parece-renesmee-de-crepusculo/

8 https://maldita.es/malditatecnologia/20230630/voces-sinteticas-impacto-positivo-recuperar-voz-enfermos-accidente/

de conversión de texto a voz. Es lo que se conoce como «clonación de voces» mediante almacenamiento de muestras de voz del paciente. También existen proyectos de investigación de ciencia ciudadana que promueven la donación de nuestra voz para crear un «banco de voces sintéticas» y proporcionar sistemas de conversión de texto a voz personalizados.[9]

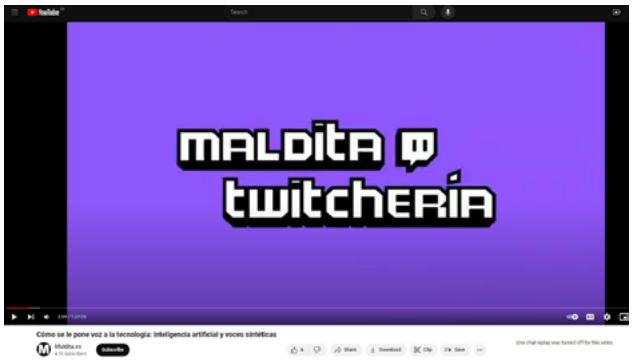

Figura 2. Canal de YouTube de Maldita.es. *Maldita Twitchería.* Pódcast recomendado: «Cómo se le pone voz a la tecnología: inteligencia artificial y voces sintéticas».

Volviendo a la cuestión de los usos consentidos y no consentidos de la IA en el ámbito de la voz, me gustaría traer a colación un caso que se encuentra precisamente a medio camino entre el consentimiento y la ausencia de este. El protagonista de este ejemplo sobre los límites de la IA en la clonación de voces es el actor estadounidense Bruce Willis, conocido por películas como *La jungla de cristal*, *El sexto sentido* o *El quinto elemento*. En 2022 anunció que padecía afasia, un trastorno del lenguaje de origen neurológico que afecta —entre otros aspectos— a la expresión oral y que, por tanto, se veía obligado a retirarse del mundo de la interpretación. Eso no impidió a la estrella de Hollywood participar —virtualmente, eso sí— en un último anuncio publicitario para un operador móvil ruso, gracias precisamente a la utilización de un *deepfake*.

Al parecer, el periódico inglés *The Telegraph* informó de que dicho *deepfake* fue posible porque el actor había vendido sus derechos de interpretación. Después se desveló que esto no era del todo cierto, gracias a la aclaración de un representante de Willis. Según leemos en este artículo de Wired,[10] la empresa que realizó el anuncio nunca contó con los derechos para el uso de la voz de Willis, sino que únicamente había llegado a un acuerdo que le permitía aplicar una versión digital de su apariencia en otro actor dentro de ese anuncio en concreto.

Figura 3. Stephen Fry. US Embassy London, https://www.flickr.com/photos/usembassylondon/27595569992/, dominio público. https://commons.wikimedia.org/w/index.php?curid=49663171

Por otro lado, tenemos el caso de Stephen Fry. Este sí que es un caso de robo en toda regla; en este caso, robo de la voz. Stephen Fry es un actor y cómico británico, cuya voz es conocida por miles de lectores en todo el mundo, ya que narró en formato de audiolibro los siete volúmenes de la saga de Harry Potter en inglés. Lógicamente, estamos hablando de muchas horas de grabación. Cualquiera que quisiera entrenar un sistema de clonación de voz mediante IA lo tendría fácil. Esto es precisamente lo que ocurrió cuando hace pocos meses apareció un documental histórico que utilizaba la voz de Stephen Fry para la narración. La sorpresa del actor fue

9 https://aholab.ehu.eus/ahomytts/

10 https://es.wired.com/articulos/deepfake-de-bruce-willis-es-problema-de-todos

mayúscula cuando lo descubrió, pues él nunca había dado su consentimiento para ese uso comercial de su voz:[11]

> It could therefore have me read anything from a call to storm parliament to hard porn, all without my knowledge and without my permission. And this, what you just heard, was done without my knowledge. So, I heard about this, I sent it to my agents on both sides of the Atlantic, and they went ballistic—they had no idea such a thing was possible.[12]

Para el que quiera conocer mejor este caso, se menciona también en este pódcast de la BBC Radio 4: *Deepfakes and the Law (Law in Action),*[13] emitido el pasado 7 de noviembre de 2023.

El uso pernicioso de los deepfakes *ha llevado a que la tecnología se convierta en un arma para acosar y humillar a las mujeres.*

El caso de Stephen Fry no difiere demasiado de lo expuesto en la primera sección de este artículo sobre las reivindicaciones del sector del doblaje y los artistas de la voz. Sin embargo, hay una diferencia esencial entre los intérpretes y dobladores anónimos, por un lado, y los cómicos y actores famosos, por otro. En ambos casos, el delito sería el mismo: se trata del hurto de una parte de la identidad de la persona. En el caso de personajes célebres, sin embargo, las repercusiones del uso ilícito de su voz pueden ser catastróficas, precisamente por la popularidad de estas personas, frente a los trabajadores de la voz que suelen vivir en el anonimato. En palabras del propio Stephen Fry, su voz podría llegar a utilizarse para llamar a asaltar el Parlamento, por poner un ejemplo. Si se emite un mensaje falso con su voz, y este se hace viral, las consecuencias políticas son inimaginables. El otro mal uso de la voz que le preocupa al actor es que se utilice para simular prácticas pornográficas por su parte. Aquí estaríamos ante un claro ejemplo de difamación de un personaje público. El porno de venganza es precisamente uno de los principales usos perniciosos de la IA, que desgraciadamente afecta principalmente a mujeres, según un informe publicado por la empresa tecnológica DeepTrace, que recoge este artículo de 2019 del MIT Technology Review.[14]

Ya a finales de 2018, cuando empecé a interesarme científicamente en los *deepfakes* y comencé a escribir mis primeras solicitudes de proyectos de investigación en este ámbito, leía con asombro y tristeza en este reportaje del *Washington Post*[15] cómo el uso de *deepfakes* en la industria del contenido pornográfico había experimentado un aumento vertiginoso. Estamos hablando de hace cinco años. Por entonces solo se oía hablar de casos aislados que afectaban sobre todo a celebridades, como Scarlett Johansson o Jennifer Lawrence. Sin embargo, la tecnología se había convertido ya en un arma para acosar y humillar a las mujeres, pues permitía generar imágenes y crear vídeos sorprendentemente realistas a partir de fotografías obtenidas fácilmente de internet.

11 https://deadline.com/2023/09/harry-potter-uk-audiobooks-narrator-stephen-fry-warns-ai-ripoff-1235548993/

12 «Por lo tanto, podría hacerme leer cualquier cosa, desde una llamada a asaltar el Parlamento hasta pornografía dura, todo sin mi conocimiento y sin mi permiso. Y esto, lo que acabas de escuchar, se hizo sin mi conocimiento. Entonces, cuando me enteré de esto, se lo envié a mis agentes a ambos lados del Atlántico y se pusieron furiosos: no tenían ni idea de que tal cosa fuera posible».

13 https://podcasts.apple.com/za/podcast/deepfakes-and-the-law/id265307843?i=1000634023260

14 https://www.technologyreview.com/2019/10/07/132735/deepfake-porn-deeptrace-legislation-california-election-disinformation/

15 https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/

Como señalan en el blog de Igualdad de Género del Banco Iberoamericano de Desarrollo,[16] los *deepfakes* se han utilizado con fines varios: como forma de extorsión, para la venta ilegal de contenido explícito, para la manipulación de discursos políticos o como una forma de burla, tanto hacia celebridades como hacia personas privadas. Efectivamente, en 2024 la manipulación de las imágenes que las mujeres suben libremente a internet no es solo un problema que afecte a las actrices famosas. Las mujeres comunes y corrientes lo sufren igualmente; y lo que es más grave todavía: las menores lo están padeciendo también. Baste recordar el caso de los falsos desnudos de las menores de Almendralejo (Badajoz), noticia que saltó a la prensa el pasado mes de septiembre de 2023. Las imágenes, difundidas por compañeros de las menores, se habrían creado con una herramienta en línea, gratuita y de fácil acceso. Así lo explica Protección Data, una empresa especializada en protección de datos y negocios digitales, en una entrada reciente de su blog.[17] En ella podemos encontrar infinidad de ejemplos concretos del uso de herramientas para la clonación de voces, desde los audios creados para hacer creer que la actriz Emma Watson recitaba pasajes del *Mein Kampf* de Adolf Hitler hasta los ejemplos que involucran a muchos otros políticos. Algunos de ellos los menciono en mi reciente libro *La fonética forense. Nuevos retos y nuevas líneas de investigación* (San Segundo, 2023a).

> *Los deepfakes se utilizan para extorsionar, para la venta ilegal de contenido explícito, para la manipulación de discursos políticos o como una forma de burla.*

Lo cierto es que, si la manipulación de la imagen afecta principalmente a mujeres, sobre todo a profesionales de la música y el cine, cuando la manipulación es de la voz y no de la imagen es habitual que la diana apunte a un tipo muy concreto de «famosos»: los políticos. Existen muestras de su voz en todas partes y al alcance de todo el mundo: no solo hablan prácticamente a diario en radio y televisión, sino que también existen discursos enteros disponibles en canales como YouTube. Para el entrenamiento de una IA generativa, es fundamental que existan datos numerosos y variados, con el fin de dotar de mayor realismo a la voz de la persona clonada.

Existen innumerables ejemplos de vídeos en los que aparecen políticos emitiendo mensajes falsos. Los motivos son diversos: desde difamar a una persona o simplemente generar *clickbait* hasta motivos serios, como crear noticias falsas con el fin de manipular a la gente: influir en elecciones o en decisiones políticas.

Uno de los ejemplos más tempranos lo presenciamos en 2018, con el *deepfake* de Barack Obama, cuya voz fue manipulada para insultar a Trump.[18] En este caso en concreto, el vídeo se realizó para hacer reflexionar al público sobre los peligros de la manipulación mediática y en el propio vídeo queda claro que Obama nunca pronunció nada de lo que aparece diciendo al principio del vídeo. Sin embargo, también encontramos casos más serios, como el considerado primer *deepfake* usado en un conflicto armado. Nos referimos a un vídeo falso del presidente de Ucrania, Zelenski, que apareció al principio del conflicto bélico con Rusia en 2022.[19] En el vídeo, que se publicó en un tabloide en lengua rusa, aparece el líder ucraniano pidiendo la rendición de sus tropas. Pese a que el medio acusó a *hackers* enemigos de crear y publicar ese *deepfake* en su web y el propio Zelenski también lo desmintió, lo cierto es que la proliferación de este tipo de vídeos ocasiona que cada vez sea más difícil

16 https://blogs.iadb.org/igualdad/es/deepfakes-violencia-basada-en-genero-inteligencia-artificial/#:%7E:text=En%20lo%20que%20respecta%20a,de%20naturaleza%20%C3%ADntima%20o%20sexual

17 https://protecciondata.es/deepfakes-y-proteccion-de-datos/

18 https://www.youtube.com/watch?v=cQ54GDm1eL0&t=1s

19 https://www.youtube.com/watch?v=X17yrEV5sl4&t=3s

distinguir las noticias reales de las falsas, o bulos (en inglés, *fake news*). De ahí que estemos presenciando una falta de confianza sin precedentes en los medios, con las consecuencias políticas y sociales que esto puede generar.

En el análisis de un posible deepfake *convendría distinguir las rasgos vocales (esto es, de la voz) de otros fenómenos que caracterizan el habla.*

¿Puede la IA alterar un proceso democrático? Que se lo pregunten a Martí Batres, gobernador de Ciudad de México. En noviembre de 2023 se hizo viral un audio de WhatsApp en el que se puede escuchar a alguien que supuestamente es Martí Batres conspirando para obstaculizar la candidatura de Omar García Harfuch, uno de los precandidatos de otro partido para las elecciones de la ciudad. El propio Martí no tardó en desmentir el supuesto bulo. Lo hizo a través de un tuit: «No soy yo, es una voz generada por inteligencia artificial». Sin embargo, el debate ya estaba servido: ¿cómo distinguir una voz real de una artificial? Ante la novedad de encontrar un supuesto caso de voz generada por IA, ahora también en español, algunos ingenieros y abogados especializados en nuevas tecnologías se pronunciaron al respecto en este artículo de *The Wired*.[20] Por suerte, existe una disciplina científica que se encarga del estudio de la voz y el habla humanas. Se llama «fonética». ¿Qué podrían decir los fonetistas al respecto? ¿En qué se fijaría un fonetista al escuchar el fragmento de audio de Batres, originalmente compartido a través de TikTok?

2.2. Unas pinceladas de ciencia

En el audio dubitado[21] de Martí Batres,[22] de apenas treinta segundos, percibimos varios fenómenos del habla y gran diversidad de características vocales (es decir, relacionadas específicamente con la voz). No solo las escuchamos, sino que, en su mayoría, las podríamos medir y analizar acústicamente.

Hay que señalar que la voz y el habla no son exactamente lo mismo. Es cierto que ambos términos se utilizan de formas diversas, a veces indistintamente. Por ello, no resulta fácil proporcionar una definición unívoca. Como indica la investigadora Jody Kreiman (2011, p. 5), «la amplia gama de funciones que cubre la voz ha dificultado la tarea de proporcionar una única definición polivalente que resulte válida y útil para todas las disciplinas, tradiciones académicas y aplicaciones científicas». En palabras del estudioso de la voz Johann Sundberg, «todo el mundo sabe qué es la voz hasta que intenta definirla». En el uso científico habitual —y el que sigue Kreiman, por cierto—, el término «voz» tiene un significado físico y una base fisiológica que se refiere a la señal acústica generada por el sistema de producción de voz del ser humano. Así, existen dos tendencias fundamentales a la hora de definir una voz. Para los que optan por una definición muy estrecha en términos fisiológicos, la voz es el sonido producido por la vibración de las cuerdas vocales. Cuando se usa en este sentido, por ejemplo, Brackett (1971),

20 https://es.wired.com/articulos/voces-generadas-por-ia-indistinguibles-las-humanas-en-espanol

21 En fonética forense, más concretamente en la tarea conocida como comparación forense de hablantes, llamamos *dubitada* a la grabación de un hablante desconocido, que comparamos con la grabación de uno o varios sospechosos (muestras indubitadas). Véase San Segundo (2023a, 2023b). La *fonética forense*, en un sentido amplio, «es la disciplina que se encarga de cualquier aplicación legal de la fonética; es decir, de aplicar los conceptos y métodos de la fonética general a la investigación y resolución de delitos en los que el habla o la voz están de algún modo implicados» (San Segundo, 2023a, p. 15).

22 https://www.youtube.com/watch?v=Z5sAjIJCKmI

«voz» se distingue claramente de «habla». Por otro lado, siguiendo siempre a Kreiman, la voz como fenómeno fisiológico y físico también se puede definir de una manera más amplia como sinónimo de habla.

En resumen, aunque los términos «voz» y «habla» a veces se usan indiscriminadamente, incluso por los propios científicos, en el análisis de un posible *deepfake* convendría distinguir las rasgos vocales (esto es, de la voz) de otros fenómenos que caracterizan el habla.

1) Características de la voz

Las características acústicas de la voz que son susceptibles de medición y análisis acústico en una grabación como la de Batres serían: la frecuencia fundamental (f_0) del hablante, la configuración espectral y la estructura formántica de sonidos concretos, así como su amplitud (dB) y tiempo (ms). Los correlatos auditivos de dichas características acústicas son el tono (agudo o grave), el timbre, la intensidad y la duración. El timbre se puede considerar sinónimo de cualidad de voz. Según autores como Kreiman, se refiere a la impresión perceptiva que se produce como resultado de una señal acústica, de forma análoga a la distinción entre «frecuencia» (una propiedad física de la vibración) y «tono» (la sensación en el oyente de dicha vibración).

2) Diversos fenómenos del habla

Existen fenómenos del habla lingüísticos, pero también paralingüísticos y extralingüísticos. Los más habituales del segundo tipo en conversaciones espontáneas son las risas, las respiraciones audibles o diversos tipos de *clics* (por ejemplo, un chasquido apicoalveolar). Para caracterizar a un hablante, en fonética forense se puede medir sencillamente la frecuencia de aparición de cada fenómeno (número de casos u ocurrencia). Asimismo, y más habitualmente, se pueden analizar acústicamente la mayoría de los fenómenos atendiendo a una o varias de las características que indicamos en el apartado anterior: tono, timbre, intensidad y/o duración.

Una división habitual que hacemos los fonetistas es aquella que distingue entre los elementos denominados segmentales —vocales y consonantes— y otros fenómenos fonéticos que afectan a más de un segmento. Estos se conocen como elementos suprasegmentales o prosódicos (véase, por ejemplo, la página web del fonetista Joaquim Llisterri).[23] Los más relevantes para el análisis de esta grabación serían:

- Melodía y entonación.

  La representación acústica de la melodía viene dada por la evolución temporal de la f_0. Cabría analizar la curva melódica de los distintos enunciados de la grabación, mayoritariamente con modalidad enunciativa, salvo la pregunta del final (*¿Me confirmas de enterado, porfa?*), que sería un enunciado con modalidad interrogativa.

- Pausas.

  Existen dos tipos de interrupciones del discurso: las pausas llenas (o sonoras) y las pausas vacías (también llamadas silenciosas). Estas últimas se manifiestan mediante un silencio, de mayor o menor duración. Generalmente se deben a la necesidad de respirar del hablante, pero también pueden tener una función demarcativa o estilística. Las pausas llenas, por su parte, suelen ser *pausas de duda* y están relacionadas con la planificación del discurso. Se suelen realizar como alargamientos vocálicos, con distintas realizaciones según la lengua. En español, lo más habitual es dudar así: «eh» (en transcripción fonética [eː]). En inglés no es raro encontrar también un elemento nasal: «ehm» [əːm]. En el audio de Batres, las pausas vacías son mínimas o muy breves. Tampoco encontramos pausas llenas. En cualquier caso, dado

23 https://joaquimllisterri.cat/

el canal (audio de WhatsApp) y la brevedad del mensaje (no se trata de un diálogo como tal), no tendría por qué llamar la atención la ausencia de ambos tipos de pausa. Habría que comparar este audio con mensajes similares enviados por el mismo hablante por este mismo medio de comunicación.

- Velocidad de articulación y diversos fenómenos rítmicos.

  Para caracterizar a un hablante, lo más habitual es que un fonetista forense mida la velocidad de articulación, o número de sílabas por segundo, excluyendo pausas y disfluencias. No obstante, existen otros enfoques rítmicos cuantitativos, por ejemplo, aquellos que miden la variabilidad de duración de intervalos vocálicos y consonánticos.

Para un listado más detallado de los parámetros fonéticos que podemos analizar en una grabación sospechosa de constituir un *deepfake*, referimos al lector a la tabla elaborada por San Segundo (2023a, p. 86), que recoge los posibles parámetros que analizan los fonetistas forenses para comparar las muestras dubitada e indubitada. Como señala el manual de buenas prácticas de la Red Europea de Institutos de Ciencias Forenses (ENFSI) y recoge también San Segundo (2023a), los parámetros que el experto opta por analizar pueden variar de un peritaje a otro, dependiendo del material disponible en las grabaciones y de qué considere importante cada experto.

Finalmente, existen otros aspectos discursivos —no necesariamente analizados desde un punto de vista acústico— en los que un fonetista se fija al analizar una grabación de voz. Por ejemplo, en la grabación dubitada de Martí Batres aparecen las siguientes voces: *Oye*, *¿no?*, *eh*, *porfa*. Se trata de unidades lingüísticas que sirven como organizadores del discurso, pueden alcanzar una gran variedad de valores semánticos, su distribución es muy versátil y a veces constituyen lo que coloquialmente conocemos como muletillas. Es habitual en lingüística forense que la elección de unos marcadores frente a otros revele la pertenencia de un hablante a un grupo concreto (sociolecto). Algunos de estos marcadores simplemente responden a la función fática del lenguaje (marcadores conversacionales de control del contacto con el interlocutor, como ese *oye*), pero la frecuencia de uso puede ser bastante idiosincrática de un hablante.

> *En el ámbito de los* deepfakes *de imagen, un* cheapfake *sería, por ejemplo, una imagen generada por IA que muestra fallos sustanciales al observarla con mayor detalle.*

Si un marcador es habitual en la variedad de español del hablante de la grabación dubitada, no será muy identificativo a nivel individual. Además, suele ser lo primero que calcan los imitadores. Si, por el contrario, una palabra o expresión es poco frecuente en la comunidad de habla a la que pertenece el hablante (variación interlocutor) y este tiende a usarla de manera constante (escasa variación intralocutor), esa palabra o expresión tendrá gran capacidad discriminatoria. Lo mismo ocurre con los rasgos vocales y los fenómenos del habla que hemos descrito anteriormente. La rareza —en el sentido de baja frecuencia de aparición— de cualquier aspecto, segmental o suprasegmental, es clave a la hora de valorar si un aspecto vocal puede realmente identificar a un hablante. Volviendo al supuesto *deepfake* de Batres, deberíamos compararlo con alguna muestra de su voz que sepamos, a ciencia cierta, que es suya y analizar cómo ambas grabaciones se asemejan o difieren para los distintos parámetros en los que un fonetista forense se fijaría.

Lo que está claro es que el audio no es un *cheapfake*. Atrás han quedado las primeras clonaciones de voz, que sonaban tan artificiales (por ejemplo, por la ausencia de marcadores o pausas realistas y por la presencia de un timbre metálico o una entonación plana). En el ámbito de los *deepfakes*

de imagen, un *cheapfake* sería, por ejemplo, una imagen generada por IA que, a simple vista, puede parecer una foto de un humano real, pero si nos paramos a observarla con mayor detalle, vemos, por ejemplo, que las manos tienen más de cinco dedos. En Maldita.es nos aconsejan cómo identificar *deepfakes* de imagen y vídeo no profesionales;[24] básicamente, debemos atender a imperfecciones y detalles, como el que acabamos de explicar, y otros, como podrían ser rostros borrosos, perspectivas erróneas, desajuste del color de piel, etc.

Sin embargo, la voz es hoy el gran reto en la detección de *deepfakes* y queda mucho por investigar. No encontramos realmente una lista de pistas a las que podamos atender para saber cuándo estamos ante un posible *deepfake* de voz, más allá de algunos consejos básicos que encontramos en páginas web como la del Incibe (Instituto Nacional de Ciberseguridad)[25] y que afectan, sobre todo, a formatos multimodales, como los vídeos, en los que hay voz, pero también imagen. Allí se recomienda básicamente que «afinemos el oído»: «Si el sonido no coincide con la imagen, detectas algún tono fuera de lugar en la voz del protagonista o una falta de sincronización, posiblemente se tratará de una falsificación». También recomiendan sospechar de los vídeos especialmente cortos, pues en ellos es más fácil camuflar los posibles errores y lograr una simulación lo más detallista posible.

Con un objetivo claramente aplicado (distinguir *deepfakes* de voces reales, evitar delitos de robo de identidad vocal, etc.) nace el proyecto de investigación del que soy investigadora principal: *¿Qué hace humana a una voz? Hacia una mejor comprensión de las características fonéticas que permiten distinguir voces reales de deepfakes* (Proyecto PID2021-124995OA-I00 financiado por MCIN/AEI/10.13039/501100011033 y por Feder. Una manera de hacer Europa).

Entre los objetivos del proyecto se encuentra descubrir científicamente qué hace humana a una voz, para así poder distinguirla de los *deepfakes*. El proyecto nace con una vocación claramente divulgativa y pretende dialogar, desde la lingüística, con otras disciplinas científicas que se ocupan también de abordar los retos de la IA en nuestra sociedad. Un ejemplo de ello es nuestra participación en la mesa redonda *Ciencia ante la desinformación: Fake news e inteligencia artificial* el pasado 8 de noviembre de 2023.[26]

Conceptos de IA relacionados con la voz

Deepfake: voz inglesa que surge de la unión de los términos *deep* (de *deep learning*, aprendizaje profundo, y *deep neural networks*, redes neuronales profundas) y *fake* (falso).
Cheapfake: *deepfakes* de baja calidad; por ejemplo, vídeos manipulados con herramientas baratas y accesibles (frente a las creaciones realizadas por un equipo de profesionales, generalmente con una mayor inversión económica). Aunque para la realización de los *cheapfakes* también se puede usar tecnología de redes neuronales, el resultado es de menor calidad y por eso suelen ser más fáciles de detectar.
Vishing: tipo de estafa de ingeniería social, similar al *phishing* y al *smishing*. En este caso, se trata de un fraude telefónico: a través de una llamada, se suplanta la identidad de una empresa, organización o persona de confianza, con el fin de obtener información personal y sensible de la víctima. El término es una combinación del inglés *voice* (voz) y *phishing*, que a su vez es un término que proviene del inglés *fishing* (pesca), pues hace alusión al uso de un cebo para que las víctimas muerdan el anzuelo.

24 https://maldita.es/malditatecnologia/20231026/detectar-deepfakes-diferencia-contenidos-inteligencia-artificial/

25 https://www.incibe.es/ciudadania/blog/deepfakes-como-se-aprovechan-de-esta-tecnologia-para-enganarnos

26 https://www.youtube.com/watch?v=T93skphUmgs

Inteligencia artificial (IA): según el Consejo Europeo,[27] es el uso de tecnología digital para crear sistemas capaces de realizar tareas que, por lo general, se considera que requieren inteligencia humana.
Aprendizaje profundo (deep learning): se trata de un tipo de aprendizaje automático que se basa en la manera en la que el cerebro humano procesa información: aprende a través de las llamadas «redes neuronales profundas». Permite procesar grandes volúmenes de datos y aprender sin ser programado explícitamente. Existen dos tipos fundamentales de aprendizaje en las redes neuronales: *aprendizaje supervisado* y *aprendizaje no supervisado.*
Entrenamiento de una red neuronal: como explica este artículo de la MIT News Office,[28] cuando se entrena una red neuronal todos sus pesos y umbrales se establecen inicialmente en valores aleatorios. Los datos de entrenamiento se envían a la capa inferior (la capa de entrada) y pasan a través de las capas sucesivas, multiplicándose y sumando de maneras complejas hasta que finalmente llegan, radicalmente transformados, a la capa de salida. Durante el entrenamiento, los pesos y umbrales se ajustan continuamente hasta que los datos de entrenamiento con las mismas etiquetas produzcan resultados similares de manera consistente.
Ataque de suplantación de identidad (*spoofing attack*): ataque malicioso realizado a un sistema biométrico para hacerse pasar por una persona autorizada, por ejemplo, en un sistema de reconocimiento de voz, obteniendo acceso no autorizado al mismo. Como explica San Segundo (2023a), existen tres tipos fundamentales: imitación, repetición (*replay*) y síntesis y/o conversión de voz.

Figura 4. Glosario de términos de IA relacionados con los *deepfakes* de voz.

3. De la ciencia al derecho: una mirada a los retos futuros

¿Cuáles son las características que hacen única la voz humana y que, por tanto, no pueden replicarse artificialmente? ¿Existe alguna característica esencial que distinga la voz humana o, por el contrario, esta será indistinguible de una voz artificial en el futuro?

Estos son los retos futuros a los que se enfrenta la fonética aplicada hoy en día, más en concreto la fonética forense, que se ocupa de las aplicaciones legales de la fonética.

En el pódcast de la BBC que recomendamos más arriba, *Deepfakes and the Law (Law in Action)*,[29] Anil Alexander, un fonetista forense que trabaja en Oxford, nos cuenta cómo están afectando los *deepfakes* a los peritajes de voz, con referencias concretas a la legislación inglesa.

Si nos centramos en el ámbito europeo, el pasado 9 de diciembre de 2023 el Consejo y el Parlamento europeos alcanzaron un acuerdo sobre el Reglamento de Inteligencia Artificial.[30] En la página web del Consejo de la Unión Europea encontramos, en varios idiomas, una explicación pormenorizada de los cambios que introduce el reglamento de IA de la UE, que podemos resumir en los siguientes puntos:

- La UE es el primer legislador mundial en intentar crear derecho sobre inteligencia artificial.
- Su propuesta legislativa puede establecer un estándar global para la regulación de la IA en otras jurisdicciones.

27 https://www.consilium.europa.eu/es/policies/artificial-intelligence/?utm_source=linkedin.com&utm_medium=social&utm_campaign=20231209-digital-ai&utm_content=visual-carousel

28 https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

29 https://podcasts.apple.com/za/podcast/deepfakes-and-the-law/id1653078437i=1000634023260

30 https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

- Similar al Reglamento General de Protección de Datos (RGPD) para la privacidad de datos, promueve el enfoque europeo en la regulación tecnológica global.

- La UE busca un enfoque ético, seguro y confiable para la IA a nivel mundial.

- El Reglamento de Inteligencia Artificial clasifica los riesgos en cuatro niveles y aplica normativas diferentes según el nivel de riesgo (véase la figura 5).

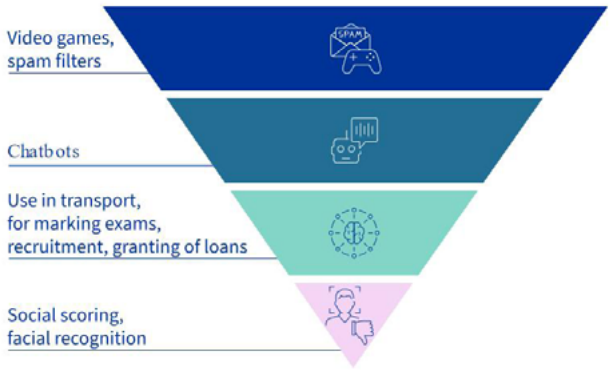

Figura 5. Diagrama que muestra una pirámide invertida compuesta por cuatro tipos de sistemas de IA que corresponden a diferentes niveles de riesgo. Este diagrama es interactivo en la web del Consejo de Europa, que permite explorar el gráfico pulsando en los ejemplos de IA para ver el nivel de riesgo y las normas correspondientes.

Me gustaría terminar con un llamamiento a más colaboraciones interdisciplinares e internacionales para la realización de proyectos de investigación que permitan detectar *deepfakes* y desarrollar sistemas robustos para la detección de ataques de suplantación de identidad. En el proyecto actual que lidero, conformado por fonetistas, logopedas e ingenieros, tratamos de aportar a esta área de conocimiento nuestra experiencia en el análisis fonético acústico-perceptual y en la comparación forense de hablantes muy similares, fundamentalmente gemelos (San Segundo, 2014). Precisamente sobre gemelos realizó también su tesis doctoral la investigadora australiana Debbie Loakes, que ahora se ha interesado científicamente por la comparación de voces reales y voces clonadas con IA. En el blog del Research Hub for Language in Forensic Evidence[31] nos deja un interesante análisis fonético —a modo de artículo divulgativo— comparando su propia voz con la clonación propuesta por un sistema de IA. Espero que los conceptos fonéticos que se explicaron en el apartado 2.2 sirvan al lector para entender qué aspectos de la voz y el habla ha analizado esta fonetista forense y, sobre todo, que esta pieza abra el apetito de mucha más gente por seguir disfrutando de esta interesante disciplina científica que es la fonética.

Bibliografía

Brackett, I. P. (1971), «Parameters of voice quality», en L. E. Travis (ed.), *Handbook of Speech Pathology and Audiology*, Nueva York, Appleton-Century-Crofts.

Kreiman, J., y Sidtis, D. (2011), *Foundations of voice studies: An interdisciplinary approach to voice production and perception*, Oxford, Wiley-Blackwell.

San Segundo Fernández, E. (2023a), *La fonética forense. Nuevos retos y nuevas líneas de investigación,*[32] Barcelona, Octaedro.

San Segundo Fernández, E. (2023b), «La fonética forense: qué es y cuáles son sus principales áreas de aplicación»,[33] *Círculo de Lingüística Aplicada a la Comunicación,* vol. 94, pp. 175-187.

San Segundo Fernández, E. (2014). *Forensic speaker comparison of Spanish twins and non-twin siblings.* Tesis doctoral, Consejo Superior de Investigaciones Científicas y Universidad Internacional Menéndez Pelayo.

31 https://blogs.unimelb.edu.au/language-forensics/2023/11/22/is-your-voice-really-your-voice-lets-ask-ai-debbie/

32 https://octaedro.com/libro/la-fonetica-forense/

33 https://revistas.ucm.es/index.php/CLAC/article/view/79972

Recursos digitales y páginas web

Pódcast *Cómo se le pone voz a la tecnología: inteligencia artificial y voces sintéticas*. https://www.youtube.com/watch?v=9wtdFQAH2zU. Canal de Maldita.es, *Maldita Twitchería.*

Pódcast de la BBC Radio 4 *Deepfakes and the Law (Law in Action),*[34] emitido el pasado 7 de noviembre de 2023.

Página web del fonetista Joaquim Llisterri, con información sobre los elementos suprasegmentales del habla. https://joaquimllisterri.cat/

Mesa redonda organizada por la Agencia Estatal de Investigación (AEI): «Ciencia ante la desinformación: *fake news* e inteligencia artificial» (8 de noviembre de 2023). https://www.youtube.com/watch?v=T93skphUmgs

Página web del proyecto de investigación *¿Qué hace humana a una voz? Hacia una mejor comprensión de las características fonéticas que permiten distinguir voces reales de deepfakes*[35] (Proyecto PID2021-124995OA-I00 financiado por MCIN/AEI/10.13039/501100011033 y por Feder. Una manera de hacer Europa).

Página web[36] de la Red Europea de Institutos de Ciencias Forenses (ENFSI: European Network of Forensic Science Institutes).

Página web del Consejo Europeo con información sobre la IA y los cuatro niveles de riesgo. https://www.consilium.europa.eu/en/policies/artificial-intelligence/?utm_source=linkedin.com&utm_medium=social&utm_campaign=20231209-digital-ai&utm_content=visual-carousel

Entrada del blog de la investigadora Debbie Loakes (Research Hub for Language in Forensic Evidence, Universidad de Melbourne, Australia): «Is your voice really your voice?». https://blogs.unimelb.edu.au/language-forensics/2023/11/22/is-your-voice-really-your-voice-lets-ask-ai-debbie/

34 https://podcasts.apple.com/za/podcast/deepfakes-and-the-law/id265307843?i=1000634023260

35 https://voicedeepfakes.github.io/

36 https://enfsi.eu/