HUMAN REVIEW | 2022 | ISSN 2695-9623
International Humanities Review / Revista Internacional de Humanidades
https://doi.org/10.37467/revhuman.v11.3967

NUEVAS TECNOLOGÍAS EN LA EDUCACIÓN SUPERIOR VIRTUAL

Las tecnologías del ultrafalso y de la traducción cara a cara

New technologies in virtual higher education
The deepfake and the face-to-face translation technologies

JOAN MIQUEL-VERGÉS [1]
[1] Universidade de Vigo, España

KEYWORDS

Deepfake
Deepfake face
Deepfake voice
Face swap
Face-to-face translation
Lip synchronization
Voice swap

ABSTRACT

In the field of virtual higher education at the University of Vigo (UVigo, Spain) there is a massive videoconference service, integrated into the Remote Campus, which can be used for virtual teaching; and, optionally, also for the automatic recording of these classes in videos. The appearance of new technologies such as deepfake and face-to-face translation will allow, in addition to the translation/dubbing of the audio in these video classes, the manipulation of the image so that the change of language is not noticed when speaking; so that lip synchronization is also done automatically.

PALABRAS CLAVE

Intercambio de caras
Intercambio de voces
Sincronización de labios
Traducción cara a cara
Ultrafalso
Ultrafalso de caras
Ultrafalso de voces

RESUMEN

En el ámbito de la educación superior virtual de la Universidade de Vigo (UVigo, España) existe un servicio de videoconferencia masiva, integrado en el Campus Remoto, que puede emplearse para la docencia virtual; y, opcionalmente, también para la grabación automática de dichas clases en vídeos. La aparición de nuevas tecnologías como la del ultrafalso y la traducción cara a cara permitirán, además de la traducción/doblaje del audio en dicha videoclases, la manipulación de la imagen para que no se note el cambio de idioma al hablar; de manera que la sincronización de labios se realice también de manera automatizada.

Recibido: 17/ 07 / 2022
Aceptado: 20/ 09 / 2022

1. Introducción

La COVID-19 ha acelerado la enseñanza en línea y la digitalización, en general, en la mayoría los centros de educación superior; incluso en aquellos donde solo se impartía docencia presencial y que ahora están ya capacitados para poder impartir docencia semipresencial o virtual, de ser necesario, como es el caso de la Universidade de Vigo (UVigo, España) (Miquel-Vergés, 2022a, 2022b). Todo ello ha acentuado la aparición de nuevas tecnologías para mejorar la docencia y los materiales docentes creados. Una de dichas tecnologías es la conocida como tecnología del deepfake; que, en español, se ha traducido como tecnología del ultrafalso ("Deepfake", 2022; Figueira & Oliveira, 2017; Suárez Iglesias, 2019; Westerlund, 2019). El término deepfake o ultrafalso, formado por las palabras en inglés deep learning (aprendizaje profundo) y fake (falso), alude, principalmente, a los sistemas informáticos que permiten, generalmente mediante programas informáticos y técnicas de inteligencia artificial (utilizando para ello algoritmos de aprendizaje no supervisados, conocidos en español como redes generativas antagónicas o RGA, y vídeos o imágenes ya existentes), desarrollar vídeos o imágenes manipulados extremadamente realistas; aunque, a veces, se aplique a los propios vídeos o imágenes así creados y se hable de los ultrafalsos o deepfakes.

Gracias a la tecnología de la inteligencia artificial (IA) hemos sido capaces de avanzar muy rápidamente en el ámbito el procesamiento de datos, imágenes y, hoy en día, también de vídeos. Todo ello con ayuda de sistemas y programas informáticos y los algoritmos adecuados. Y, obviamente, sin tener que manejar presupuestos económicos, ni materiales, ni recursos humanos desorbitados como pudiera ocurrir en el pasado (García Nieto, 2020; López Lafuente, 2020).

Si, por ejemplo, tomamos en consideración el primer largometraje de animación *Snow White and the Seven Dwarfs* (conocido como Blanca Nieves y los siete enanos en Hispanoamérica y Blancanieves y los siete enanitos en España) producido por Walt Disney en 1937, de 83 minutos de duración, su coste temporal, material, de recursos humanos y económico fue, a todas luces, desorbitado:

> La inversión de 500 000 dólares terminó transformándose a lo largo de los tres años de producción en un millón y medio, y el proyecto integró un equipo de 600 dibujantes, veinte directores —David Hand fue el supervisor del trabajo de estos más que un director propiamente dicho—, veintidós animadores y la friolera de dos millones de dibujos de los que sólo 400 000 fueron utilizados en el producto final. (Benítez, 2022)

Si nos ceñimos exclusivamente a su coste económico, el presupuesto destinado fue, exactamente, de 1 488 423 dólares estadounidenses de 1937 ("Snow White and the Seven Dwarfs", 2022) que, actualizados a 2022[1] en función de la inflación, equivaldrían a unos 30 207 952 dólares; es decir, unos 29 373 482[2] euros (a día 23 de julio de 2022). Ello representa, ni más ni menos, que un coste aproximado actual de unos 350 000 euros por minuto de grabación.

Hemos considerado necesario introducir de entrada estos datos relacionados con el coste por minuto de grabación para hacer patente que, si uno dispone de recursos suficientes para poder producir un vídeo, a pesar de las posibles limitaciones técnicas que puedan existir en la elaboración de este, seguramente el resultado final llegará a ser el óptimo esperado.

En el caso de las tecnologías que vamos a analizar a continuación también tomaremos en consideración este conjunto de factores que inciden en el coste del producto final; que, en nuestro caso y de manera generalizada, será la creación y traducción/doblaje automático de vídeos educativos en el ámbito de la educación superior (universitaria) virtual. Pero, a diferencia del ejemplo anterior, nuestro propósito será conseguir automatizar todo el proceso, con un coste económico (una vez implementadas las tecnologías) cero.

A continuación, estableceremos los principales objetivos de nuestro trabajo de investigación y la metodología que seguiremos para nuestro análisis y para la elección de la muestra de materiales (ejemplos) a analizar. En primer lugar, realizaremos el análisis a nivel global; para, a continuación, contextualizarlo en el ámbito específico del Campus Remoto de la UVigo.

2. Objetivos

Si bien es cierto que inicialmente el uso de los ultrafalsos o *deepfakes* fue (y sigue siendo) muchas veces con intenciones fraudulentas, en la actualidad su uso se ha extendido a otros ámbitos más beneficiosos para la sociedad como puedan ser, entre otros, su uso en el ámbito cinematográfico (p.ej., cuando un actor o actriz se queda sin voz; Dickson, 2018; Marr, 2019); en el ámbito social (p.ej., para ayudar a personas con Alzheimer o a personas que han perdido su voz; Ebert, 2009); en el ámbito empresarial (p.ej., en la campaña de la cerveza Cruzcampo que convierte a la cantante, actriz y bailadora Lola Flores en su protagonista, 25 años después de su fallecimiento; Lacort & Tones, 2021); en el ámbito histórico (p.ej., para recrear personajes históricos; The Dalí Museum, 2019; Vicomtech, 2022a, 2022b, 2022c); en el ámbito lúdico de los videojuegos (p.ej., para aumentar

1 CPI Inflation Calculator, accesible a través de la URL https://cutt.ly/DL5UtO9

2 Currency Converter (USD-EUR), accesible a través de la URL https://cutt.ly/UL5Io2G

su realismo; Nelson, 2020); o, por último, en el ámbito educativo (p.ej., para la traducción/doblaje automático de materiales digitales multimedia; Ballantine, 2019; Brown, 2019; Kuper, 2018; Solsman, 2019).

Según Pérez Navarro et al. (2020): "los vídeos digitales están cada vez más presentes en los procesos de aprendizaje, especialmente en el contexto educativo provocado por la COVID-19". Pero, "la creación de vídeos es una actividad que requiere mucho tiempo para los profesores, que generalmente no son expertos en esta tarea. Para conseguir unos cuantos minutos de vídeo, deben dedicar muchas horas".

En el ámbito de la educación superior (universitaria) virtual de la UVigo existe en la actualidad un servicio de videoconferencia masiva, integrado en el denominado Campus Remoto[3], que puede emplearse para la docencia no presencial (virtual); y, opcionalmente, puede usarse también para la grabación (automática) de dichas clases. Dicho Campus Remoto está generado sobre una plataforma propia basada en código abierto *Big Blue Button* (BBB) que, en el año 2021, sumaba ya 2603 aulas y 2895 despachos virtuales creados, en los que se realizaron casi 17 000 grabaciones de clases, vídeos educativos y actos académicos (La Región, 2021). A partir de dichos datos podemos preguntarnos si, en vez de crear nuevos vídeos educativos, no es mejor reaprovechar los vídeos educativos ya creados de manera automática en dicho Campus Remoto.

Cuando hablamos de reaprovechar los vídeos grabados en las clases virtuales del Campus Remoto nos estamos refiriendo al hecho que no es posible utilizar directamente los vídeos grabados como material docente posterior si queremos que su eficacia sea la mayor posible. Si tenemos en cuenta que la duración de una clase docente en la FFT-UVigo es de, aproximadamente, 120 minutos; difícilmente un vídeo educativo de 120 minutos puede ser directamente asimilado *a posteriori* por el alumnado de una determinada materia universitaria. Sobre la longitud óptima de un vídeo educativo, los especialistas en este ámbito remiten a la conocida como "regla 2-6-10" (López Ardao, 2017): 1) Si decidimos usar un pequeño vídeo para explicar un concepto o realizar una pequeña presentación, el vídeo puede ser ampliado sin pérdida de enganche (*engagement*) hasta los 2 minutos; 2) Más allá de los 2 minutos, el interés del vídeo se centra ya en el contenido, como es el caso de los vídeos educativos, donde el contenido prevalece. En estos casos, la duración del vídeo puede prolongarse hasta los 6 minutos; 3) A partir de ahí, cada minuto adicional constituye una pérdida de enganche y deberíamos fijar un límite máximo de 10 minutos. La recomendación a este respecto sería centrarnos en la postproducción de la clase grabada para segmentar los videos en trozos de menos de 6 minutos, a ser posible.

A este respecto, la cantidad ingente de vídeos educativos que podemos obtener *a posteriori* (con garantías educativas de longitud óptima, etc.) a partir de las grabaciones automáticas de las clases virtuales del Campus Remoto[4]; hace que el factor tiempo resulte determinante en este ámbito. Si, al igual como sucede con la generación de los vídeos en el Campus Remoto, no es posible automatizar también la edición posterior de los vídeos, hoy por hoy, resulta muy complicado poder disponer de dichos vídeos como material educativo con un mínimo de calidad. Y la perspectiva es todavía más pesimista si lo que nos proponemos es adaptar dichos vídeos (audio y vídeo propiamente dicho) a otros idiomas; donde, hoy por hoy, la automatización del proceso está todavía en fase de investigación.

"En la actualidad encontramos ya, por separado o combinadas, tecnologías que nos permiten traducir texto. También transcribir audio a texto o generar audio a partir de texto (López Lafuente, 2020)" con plenas garantías. Es en el ámbito de cómo podemos integrar ese audio en un vídeo donde todavía resta todavía mucho por investigar e implementar. Sin embargo, han empezado a aparecer ya tecnologías y programas informáticos que lo permiten, de manera automatizada y con unos mínimos aceptables de calidad en el producto final. Así, por ejemplo, Prajwal et al. (2019a, 2019b) y Rudrabha Mukhopadhyay (2019) nos proponen una nueva tecnología para la manipulación de vídeos empleando inteligencia artificial, cuyo propósito es traducir/doblar vídeos en los que habla una persona (que es la situación prototípica que nos encontraremos en los vídeos del Campus Remoto). "A partir de un vídeo cualquiera, la herramienta informática en cuestión traduce el audio y, además, manipula la imagen para que no se note el cambio de idioma al hablar (López Lafuente, 2020)"; de manera que la sincronización de labios se realiza de manera automatizada. En su proyecto, Prajwal et al. (2019b) trabajan con vídeos en inglés para traducirlos al hindi, pero, *a priori*, podría aplicarse dicha tecnología a cualquier otro idioma y tipo de vídeo; y extender su uso a otros ámbitos como podrían ser el de traducir programas informativos, traducir documentales, traducir eventos deportivos, o, traducir ficciones audiovisuales como series y películas (García Nieto, 2020; López Lafuente, 2020; Prajwal et al., 2019b). Dicha tecnología ha sido bautizada como *face-to-face translation*, que, en español, sería traducción cara a cara (Suárez Iglesias, 2019). Para entender mejor cómo funciona todo el proceso puede consultarse el vídeo de demostración creado *exprofeso* por los propios investigadores y disponible en Rudrabha Mukhopadhyay (2019).

De acuerdo con los investigadores "si bien el software no es lo suficientemente rápido para hacer esta traducción en tiempo real, LipGAN [que es el nombre del componente del módulo traductor cara a cara utilizado para generar

3 Accesible a través de la URL https://campusremotouvigo.gal

4 Obviamente, esta información magnifica todavía más el tiempo que un docente universitario debería dedicar a la creación y/o edición de sus vídeos educativos. En el caso de la UVigo, si tenemos en cuenta que el total de horas docentes de una materia de 6 créditos es de 48 horas y que las clases docentes son, mayoritariamente, de 2 horas; ello querría decir que se generan 24 vídeos de 2 horas en cada materia impartida en el Campus Remoto. Si fraccionamos dichos vídeos en segmentos de 6 minutos de duración tendríamos, al final, la friolera de 144 vídeos.

rostros realistas a partir del audio traducido, con los labios sincronizados] tiene potencial para el doblaje de películas, de vídeos educativos y para entrevistas y noticias en la televisión. 'Por ejemplo, un espectador hindi o alemán podría ver una entrevista a Obama en su idioma con su propia voz y con los labios sincronizados', afirman los autores del estudio" (García Nieto, 2020).

En nuestro caso, como ya comentamos, hoy por hoy no sería necesario que nuestro sistema tradujese los ultrafalsos educativos en tiempo real, por lo que podríamos usarlo de manera asíncrona. Sin embargo, como afirma el propio K. R. Prajwal, autor principal del estudio:

> A medida que la comunicación digital de hoy en día se vuelve cada vez más visual, argumentamos que existe la necesidad de sistemas que puedan traducir automáticamente un vídeo de una persona que habla en el idioma A a un idioma B con una sincronización labial realista. (García Nieto, 2020)

Así, pues, el objetivo de este trabajo es, por una parte, analizar si la tecnología del ultrafalso o *deepfake* y la tecnología de la traducción cara a cara o *face-to-face translation* pueden utilizarse en el ámbito de los vídeos educativos utilizados en la educación superior (universitaria) virtual, en general. Y, por otra parte, y de manera particular, analizar si es posible integrar dicha tecnología en el Campus Remoto de la UVigo; de manera que dicho entorno telemático, además de permitir grabar las clases (y conferencias, cursos de formación, etc.) de manera automática, permita también (también de manera automática) traducir/doblar los vídeos educativos generados en dicha plataforma a otros idiomas. Todo ello, además, dentro de un sistema automatizado y de coste cero. Se trataría de que el personal docente de la UVigo, que en este momento ya puede grabar sus clases docentes de manera automatizada, pueda también traducir o realizar el doblaje[5] (Real Academia Española, s.f.) de dichos videos a otros idiomas diferentes a los de los vídeos originales. Y, todo ello, de manera que el producto (vídeo) resultante resulte "natural"; esto es, de manera que el proceso de ajuste labial (y/o facial) entre las nuevas caras y voces obtenidas al final de todo el proceso de traducción/doblaje sea óptimo. Se trata de conseguir la mayor naturalidad posible en el discurso resultante.

Dado que estas tecnologías son de reciente aparición y están todavía muy poco implementadas en la sociedad y, por tanto, son poco conocidas, antes de analizarlas en profundidad realizaremos una pequeña introducción a las mismas.

3. Metodología

Para llevar a cabo nuestro análisis, en primer lugar, definiremos y explicaremos en qué consiste la tecnología del ultrafalso o *deepfake* y la tecnología de la traducción cara a cara o *face-to-face translation*. Para ello, especificaremos cuáles fueron sus antecedentes en el ámbito de las voces y las imágenes fijas de caras; esto es, la tecnología de intercambio o clonación de voces (o *voice swap*) y la tecnología de intercambio o clonación de caras (o *face swap*), respectivamente.

Con la aparición de los vídeos (integrados por multitud de imágenes fijas o fotogramas en movimiento y por audio) dichas tecnologías primigenias se mostraron inoperantes; ya que, por una parte, suponían trabajar con multitud de caras (de fotogramas), y, por otra parte, era necesario realizar un proceso de ajuste labial (y/o facial) entre las nuevas caras y voces obtenidas al final de todo el proceso, en aras a conseguir la mayor naturalidad posible en el discurso resultante. Si bien se conseguían resultados aceptables en términos de calidad, ello era casi siempre mediante costosos procesos manuales de pre y posedición.

La aparición de la inteligencia artificial y del aprendizaje profundo o *deep learning* supuso la aparición de nuevas tecnologías, como es el caso del ultrafalso de voces (clonación o intercambio de voces usando aprendizaje profundo o *deepfake voice*) y el ultrafalso de caras (clonación o intercambio de caras usando aprendizaje profundo o *deepfake face*), que agilizaron y automatizaron el proceso de creación de nuevas voces y caras, respectivamente. Sin embargo, ello no supuso tampoco cambio alguno en el proceso de ajuste labial (y/o facial); que siguió siendo manual y costoso.

A continuación, analizaremos la tecnología de la traducción cara a cara o *face-to-face translation*, que integra las tecnologías del ultrafalso de voces y de caras; y que, además, es capaz de realizar el proceso de ajuste labial (y/o facial) de manera automática y con unos costes asequibles. Constataremos que, si bien hoy por hoy los resultados obtenidos mediante el uso de dicha tecnología no permiten emplearla en tiempo real; sí permiten una automatización del proceso y unos resultados aceptables en términos de calidad, siempre y cuando se disponga de tiempo suficiente para poder llevar a cabo todo el proceso.

Por último, nos preguntaremos si es posible adaptar dicha tecnología de la traducción cara a cara al Campus Remoto de la UVigo. Dado que los programas informáticos y librerías necesarios para poder implementar dicha tecnología son de código abierto, y, dado que también el propio Campus Remoto está generado sobre una plataforma propia basada en código abierto *Big Blue Button* (BBB); inferiremos que solo es cuestión de tiempo (siempre y cuando se abogue por ello) que la tecnología de la traducción cara a cara se incorpore como una

5 Un doblaje, en cine y televisión, es la operación en la que se sustituye la voz original de un actor por otra, en distinto idioma o en el mismo.

herramienta más en el Campus Remoto de la UVigo. Por nuestra parte ya hemos contactado con los responsables de dicha plataforma de teledocencia para que evalúen la posibilidad real de hacerlo.

Para cerciorarnos de que las tecnologías del ultrafalso y de la traducción cara a cara son operativas y pueden usarse en el ámbito de la traducción automática (entendida esta como el la traducción/doblaje automático con sincronía labial incluida) de vídeos educativos, veremos algunos ejemplos prácticos donde quede patente esta agilización y perfeccionamiento del proceso de ajuste labial (y/o facial) que proporciona esta nueva tecnología en comparación con el proceso y tecnologías que se han venido usando tradicionalmente hasta ahora. Si bien estos ejemplos no son del ámbito de la educación superior (universitaria) virtual, entendemos que son lo suficientemente ilustrativos para poder apreciar el potencial de estas dos tecnologías, en general, y la posibilidad de implementarlas en el ámbito del Campus Remoto de la UVigo, en particular. Como ya hemos indicado anteriormente, a pesar de que el proceso a realizar se ha automatizado y los resultados obtenidos son cada vez más satisfactorios, todavía se requieren unos recursos y una labor de pre o posedición manual considerables si se quiere optimizar la calidad; es por ello por lo que no hemos optado por realizar nuestros propios ejemplos ilustrativos de ultrafalsos y de traducciones/doblajes cara a cara.

Los dos ejemplos de ultrafalsos integrales (de voz y de audio) que analizaremos tienen una duración aproximada de 1 minuto y poseen una calidad óptima; pero, aun así, su realización ha supuesto unos costes temporales, materiales, de recursos humanos y económicos considerables. Sobre todo, si lo comparamos con el ámbito específico del Campus Remoto de la UVigo, donde estaríamos hablando de la creación y traducción/doblaje automatizado de unos 144 vídeos de 6 minutos de duración por cada materia de 6 créditos (48 horas) impartida. La comparación entre ambas tipologías de ejemplos, hoy por hoy, no parece la más adecuada en términos de costes; pero si en términos de "potencialidad" y "extrapolabilidad", que es lo que pretendemos en nuestro caso.

3.1. La edición de vídeos

En el ámbito audiovisual, los vídeos digitales que se usan en la actualidad, en la práctica, además del vídeo propiamente dicho, incluyen también audio digital. Por lo tanto, *a priori*, editar un vídeo digital implica (o puede implicar) tener que editar, a la vez, vídeo y audio. A su vez, el vídeo propiamente dicho está compuesto también por un conjunto de imágenes fijas denominadas fotogramas:

> Un fotograma es una de las muchas imágenes fijas que componen la imagen en movimiento o video ya sea de forma digital o analógica. Desde el comienzo de la cinematografía moderna hacia finales del siglo XX, y en algunos casos en la actualidad, las imágenes individuales se grababan en una tira de película fotográfica que eran proyectadas mediante un cinematógrafo.
>
> Cuando se visualiza la imagen en movimiento (vídeo), cada fotograma aparece en la pantalla durante un breve período de tiempo (en la actualidad, 1/24, 1/25 o 1/30 de segundo) y luego se sustituye inmediatamente por el siguiente. La persistencia de la visión mezcla los fotogramas, produciendo la ilusión de una imagen en movimiento. (Bmotion, 2022)

Por lo tanto, la edición de vídeos (en su acepción más general) puede considerarse, *a priori*, como la edición de múltiples imágenes fijas (o fotogramas) y/o la edición de audio. Llegados a este punto, si tenemos en cuenta que la tasa de fotogramas estándar, expresada como fotogramas por segundo (sigla FPS, símbolo f/s, también conocida como "cuadros por segundo") ("Fotogramas por segundo", 2022), es de 24 fotogramas por segundo; podemos inferir que, en un vídeo educativo estándar de 6 minutos tendríamos alrededor de 8 640 fotogramas (o imágenes estáticas). Si editar un vídeo digital de 6 minutos supone tener que editar 8 640 imágenes estáticas, el trabajo a realizar requiere una cantidad de tiempo y esfuerzo considerable. Lo mismo sucede si hablamos de traducir/doblar un vídeo.

3.2. La clonación o intercambio de voces y de caras

Dentro de la edición de vídeo y de imágenes, la clonación, intercambio, remodelación o edición de voces o de caras mediante el uso de métodos tradicionales se conoce en inglés, tradicionalmente, con el término de *voice swap* o *face swap*, respectivamente. Mientras que, el término de ultrafalso de voces o ultrafalso de caras (que en inglés se conoce como *deepfake voice* o *deepfake face*, respectivamente), se reserva para aquellos casos en los que se usan métodos de inteligencia artificial como es el caso del aprendizaje profundo (o *deep learning*). Ambas tecnologías han despertado gran interés en la sociedad, en general, en los últimos tiempos; no solo en ámbitos especializados. Sobre todo, cuando empezó a aplicarse en el ámbito de los vídeos (donde se combina la tecnología del ultrafalso de voz y la del ultrafalso de cara), dando lugar a la tecnología del ultrafalso (a secas) o *deepfake*.

Son múltiples y variados los ámbitos de especialización donde estos diferentes tipos de tecnologías se han ido desarrollado. En el caso de la tecnología del ultrafalso o *deepfake*, aplicada a los vídeos, su conocimiento por parte

del gran público es relativamente nuevo y obedece, paradójicamente, a la aparición en 2017, en algunos hilos de usuarios en Reddit[6] (Cole, 2017), casi siempre poniendo caras de famosas a vídeos pornográficos.

3.2.1. La clonación o intercambio de voces

La clonación o intercambio de voces mediante métodos tradicionales (*voice swap*) o mediante nuevos métodos de la inteligencia artificial como el aprendizaje profundo (ultrafalso de voz o *deepfake voice*), al no tener asociadas imágenes, no presenta, *a priori*, las mismas dificultades técnicas que pueda presentar la clonación o intercambio de caras mediante métodos tradicionales (*face swap*) o mediante nuevos métodos de la inteligencia artificial como el aprendizaje profundo (ultrafalso de cara o *deepfake face*), que analizaremos más adelante. Hoy en día son muchas las aplicaciones que nos permiten "intercambiar, suplantar, o imitar" las voces de personas y personajes conocidos a partir de los audios de nuestra propia voz; e, incluso, hay ya dispositivos, como Alexa, que nos permiten integrar funciones de clonación o intercambio de voces para recuperar voces de seres queridos ya fallecidos (AWS Events, 2022; Harding, 2022).

> After noting the large amount of lives lost during the pandemic, Prasad played a video demo, where a child asks Alexa, "Can grandma finish reading me Wizard of Oz?" Alexa responds, "Okay," in her typical effeminate, robotic voice. But next, the voice of the child's grandma comes out of the speaker to read L. Frank Baum's tale. (Harding, 2022)

Alexa no es el primer producto de consumo que usa el intercambio de voces para reemplazar a un miembro de la familia que no puede estar allí en persona. El altavoz inteligente Takara Tomi, como señala Ion (2022), utiliza inteligencia artificial para leer cuentos infantiles antes de dormir con la voz de los padres. Según la información técnica del producto, los padres cargan sus voces, por así decirlo, leyendo un guion durante unos 15 minutos.

En cualquier caso, en la actualidad, cuando hablamos de la clonación o intercambio de voces no se trata (o no se trata solo) de imitar la voz en un audio y substituir el audio de la otra persona; es decir, no se trata (o no se trata solo) de enmarcar el problema como una tarea de conversión de voz, propiamente dicha. Sino que, por el contrario, en la actualidad se pretende enmarcar el problema como una tarea de generación de voz. Se trata, por tanto, de crear aplicaciones y dispositivos que permitan aprender a producir una voz de alta calidad con, por ejemplo, menos de un minuto de grabación versus horas de grabación en un estudio, que es lo que se estaba haciendo hasta ahora. Este salto cuantitativo y cualitativo se ha logrado (igual que veremos ha sucedido en el caso de clonación o intercambio de caras) con el uso de la inteligencia artificial y del aprendizaje profundo (o *deep learning*).

Como anticipo a esta nueva tecnología del ultrafalso de la cual hablaremos más adelante, comentar que existen ya en Internet plataformas gratuitas como FakeYou[7] (s.f.) que permiten convertir los textos que escribamos en un campo de textos en clips de audio, basándose en las voces de personajes famosos (reales o ficticios, vivos o muertos) que elijamos. Tiene soporte para una variedad de voces de famosos y celebridades organizados a través de diferentes categorías. Como curiosidad, a continuación, se puede acceder en la Tabla 1 a los audios de un fragmento de este mismo párrafo "hablados" por diferentes personajes (humoristas, actores, actrices), vivos y muertos; todos ellos con una voz muy reconocible para los usuarios españoles. Para poder identificarlos y poder comparar sus voces originales con sus voces intercambiadas se han insertado sendos enlaces a la Wikipedia, a su voz original en YouTube y a sus voces generadas con la aplicación en línea FakeYou (s.f.). Hay que indicar que no ha habido ningún tipo de pre o posedición en los audios; y que los mismos se han generado de manera totalmente automática por la aplicación, en cuestión de segundos.

6 Los vídeos ultrafalsos o *deepfakes* adquirieron ese nombre gracias a un usuario de Reddit (uno de los foros más populares de Internet en esa época) llamado "deepfakes" que, en el año 2017, publicó una serie de vídeos falsos pornográficos de famosas que llamaron poderosamente la atención. Samantha Cole, una reputada periodista de la revista y del canal de vídeo en línea Motherboard-VICE, (dedicados a la tecnología, la ciencia y a los seres humanos), publicó un artículo sobre el porno manipulado por la inteligencia artificial, que se centraba en un vídeo "deepfake" pornográfico en el cual aparecía la actriz Gal Gadot, de *Wonder Woman*.

7 FakeYou.com era conocida anteriormente como Vocodes; y, curiosamente, durante la redacción de este trabajo, cambió de nombre. Como ellos mismos indican en su web "FakeYou.com se conocía anteriormente como https://vo.codes, pero se le cambió el nombre para reflejar mejor su propósito y dirección. (De todos modos, no mucha gente sabe qué es un vocoder, por lo que no tenía sentido mantener el nombre)."

Tabla 1. Ejemplos de clonación o de intercambio de voces

	Wikipedia	Voz original de YouTube	Voz clonada de FakeYou
Bart Simpson	https://cutt.ly/fK225ob	https://cutt.ly/AK23d27	https://cutt.ly/vK20Y5w
Chiquito de la Calzada	https://cutt.ly/dK29qmu	https://cutt.ly/2K23zfm	https://cutt.ly/iK20PJ4
José Coronado	https://cutt.ly/HK29yCY	https://cutt.ly/ZK23Eki	https://cutt.ly/BK20DqJ
Lola Flores	https://cutt.ly/PK29oOj	https://cutt.ly/KK23DVR	https://cutt.ly/YK20FXU
Miguel Gila	https://cutt.ly/uK29aFz	https://cutt.ly/PK23NPQ	https://cutt.ly/KK20H5g

Fuente(s): Wikipedia, YouTube, FakeYou, 2022.

Frente a la clonación o intercambio de voces tradicional (o *voice swap*), el ultrafalso de voz (también denominado clonación o intercambio de voces mediante aprendizaje profundo o *deepfake voice*) supone ir un paso más allá en el proceso de replicar o generar un clon de la voz de una determinada persona o personas, al usar la inteligencia artificial y el aprendizaje profundo para llevar a cabo dicho proceso (Veritone, 2022). La tecnología ha avanzado hasta el punto de que hoy en día puede replicarse o clonarse una voz humana con gran precisión en el tono y en la semejanza de las voces.

Para poder clonar la voz de alguien es necesario disponer de datos de entrenamiento para alimentar los modelos de inteligencia artificial. Estos datos suelen ser grabaciones originales de la voz a clonar que proporcionan un excelente ejemplo de la voz de la persona objeto de análisis. La IA puede usar estos datos para generar una voz que suene auténtica, que luego se puede usar para hablar cualquier cosa que se escriba en formato de texto (esto es, la conversión de texto a voz); o bien, directamente, hablando (conversión de voz a voz).

El fin último de la clonación o intercambio de voces, tanto si es mediante métodos tradicionales (o *voice swap*) o mediante la tecnología del ultrafalso (o *deepfake voice*) es el mismo: conseguir que un texto o un audio con la voz de una determinada persona pase a ser percibido como si fuese hablado por otra persona diferente. Lo único que cambia en cada caso es la tecnología a emplear; y, por ende, el tiempo empleado para conseguir los resultados finales y los propios resultados obtenidos. Aunque se aspira a conseguir un proceso lo más automático posible, en la actualidad todavía es necesario realizar algunas labores de "retoque manual" en el producto final. A continuación, en la Tabla 2, se muestran algunos ejemplos de ultrafalsos de voz; el primero de los cuales, el de Lola Flores, forma parte de un vídeo.

Tabla 2. Ejemplos de ultrafalsos de voz

	Wikipedia	**Voz original de YouTube**	**Ultrafalso de voz**
Lola Flores	https://cutt.ly/PK29oOj	https://cutt.ly/KK23DVR	https://cutt.ly/6K8fQUO
Francisco Franco	https://cutt.ly/yK8sQQT	https://cutt.ly/FK8s9bn	https://cutt.ly/kK69GF1

Fuente(s): Wikipedia, YouTube, Vicomtech, 2022.

Estos ejemplos nos aproximan cada vez más a nuestro objetivo final, que es el de poder convertir texto a voz o voz a voz, con la voz del docente de una materia. En cuanto a calidad del producto resultante hay que destacar que, si bien los resultados son muy notables en cuanto a calidad, el tiempo empleado para su realización dista todavía mucho de lo que resultaría óptimo para nuestros interese; a no ser, claro está, que se disponga ya del *vocoder* o codificador de voz adecuado construido previamente, como sucede en los ejemplos de la Tabla 1 mostrados en la página anterior.

3.2.2. La clonación o intercambio de caras

La posibilidad de clonar o intercambiar caras de manera tradicional (o *face swap*) es una de las características que, con la aparición de los vídeos digitales, la edición de vídeos digitales y los programas para ordenadores y las aplicaciones para teléfonos móviles, ha ido adquiriendo cada vez más protagonismo. La aplicación Snapchat es uno de los ejemplos de programas (o aplicaciones) de "mensajería para teléfonos inteligentes con soporte multimedia de imagen, vídeo y filtros para fotos de realidad aumentada ("Snapchat", 2022)" que, entre sus características más populares, incorpora la susodicha posibilidad de intercambiar caras. "Actualmente, Snapchat se encuentra entre las aplicaciones más utilizadas a nivel mundial junto a Facebook, WhatsApp, Twitter, Instagram y Tik Tok ("Snapchat", 2022)"; de manera que, en el 1º trimestre de 2022, Snapchat tenía ya 332 millones de usuarios activos a diario (Costa, 2022).

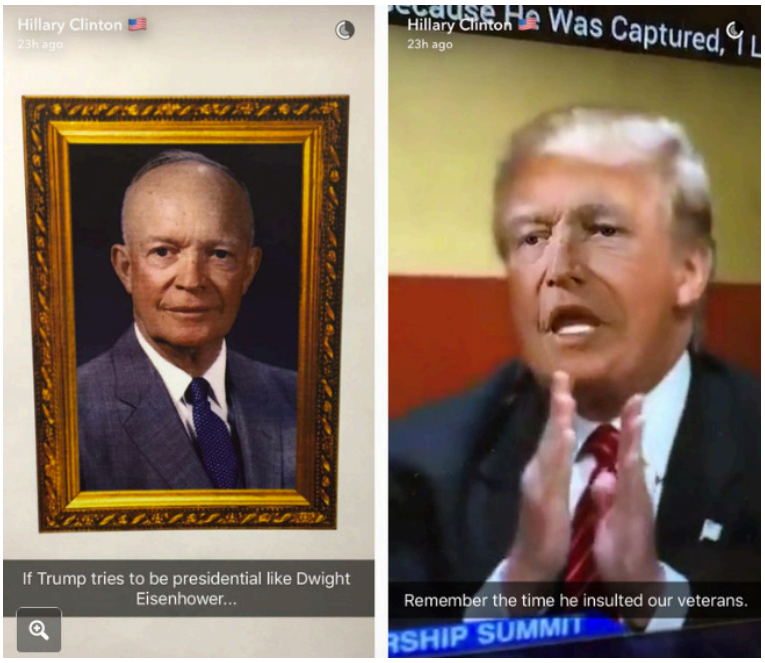
Esta posibilidad de intercambio de caras es cada vez más usada por las marcas y por los servicios de mercadotecnia para hacer propaganda de sus productos; ya sean estos productos comerciales propiamente dichos

o bien personas (Olivas, 2016). Así, por ejemplo, el servicio de *streaming* de vídeo Netflix, en Francia, colocó en la ciudad de París decenas de carteles con imágenes de los personajes de sus series originales, como *House of Cards* y *Orange is the New Black*, para que los parisinos (en particular) y los turistas (en general) hiciesen intercambio de caras o *face swap* en Snapchat (KULTT, 2016); tal como puede observarse en la Figura 1 de la página siguiente.

Figura 1. Uso del intercambio de caras tradicional o face swap para campañas de promoción de productos

Fuente(s): Olivas, 2016.

Figura 2. Uso del intercambio de caras tradicional o face swap para campañas políticas



Fuente(s): Olivas, 2016.

O, en los Estados Unidos de América, en la campaña presidencial de Hillary Clinton de 2016 que la enfrentó al candidato republicano Donald Trump, en la cuenta oficial de Hillary Clinton en Snapchat aparecían una serie de imágenes donde se hizo intercambio de caras o *face swap* de Donald Trump con expresidentes de los Estados Unidos; tal como puede observarse en la Figura 2, en este caso con la cara del expresidente D. D. Eisenhower.

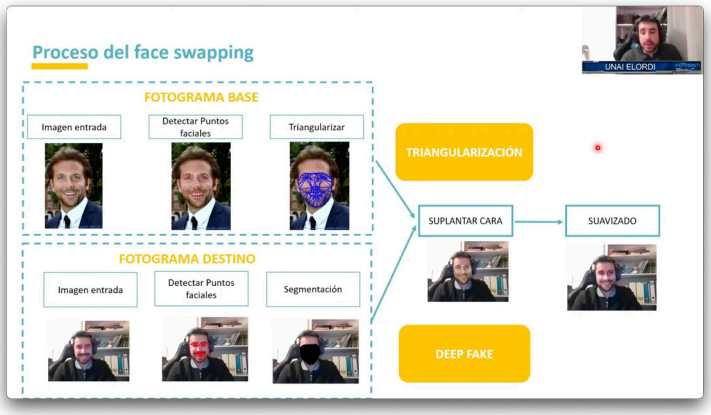
O, finalmente, el o intercambio de caras o *face swap* también ha sido utilizado en campañas de concienciación social. Así, en Brasil, National Geographic y Young & Rubicam crearon una serie de piezas que invitan a cambiar el prejuicio por el conocimiento; una de las cuales aparece en la Figura 3.

Figura 3. Uso del intercambio de caras tradicional o face swap para campañas de concienciación social

Fuente(s): Olivas, 2016.

Igual que ocurrió en el caso de la clonación o intercambio de voces tradicional (o *voice swap*) donde la tecnología del ultrafalso de voz (o *deepfake voice*) constituyó un gran avance en la tecnología al usar la inteligencia artificial, y, más concretamente, el aprendizaje profundo; también sucedió lo mismo en el caso de la clonación o intercambio de caras tradicional (o *face swap*) frente a la nueva tecnología del ultrafalso de caras (o *deepfake face*). Perov et al. (2020) y Euskal Encouter (2020) analizan las características de ambas tecnologías, qué tienen en común, en qué se diferencian, y como se complementan (Figura 4). Perov et al. (2020) es, además, el germen de DeepFaceLab (2022), un sistema de código abierto, fácil de usar, que permite conseguir unos resultados fotorrealistas de intercambio de rostros sin tener que llevar a cabo costosos ajustes.

Figura 4. Características del proceso de face swapping o de intercambio de caras



Fuente(s): Euskal Encouter, 2020, 06m59s (basado en Perov et al., 2020).

En el momento de tener que suplantar la cara es cuando podemos optar por dos tipos de tecnología: una más tradicional, que es la "triangulación" (Figura 5), en la cual obtenemos la textura de nuestra cara base, lo dividimos

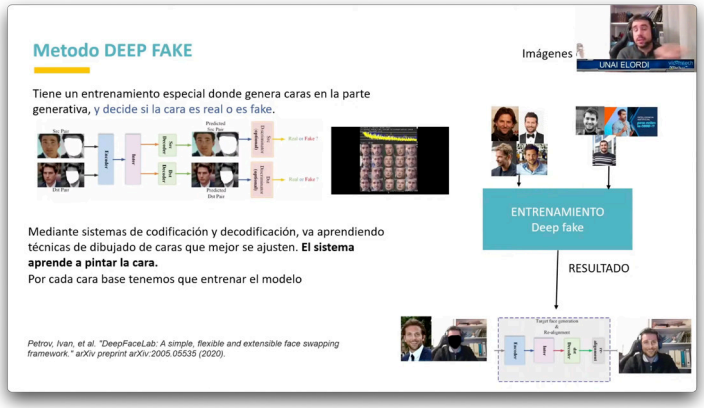
en triángulos, y lo deformamos para que se ajuste a la cara de destino (este proceso lo podemos realizar de manera manual, semiautomática o automática).

Figura 5. Método o tecnología de "triangulación" tradicional para el intercambio de caras

Fuente(s): Euskal Encouter, 2020, 18m45s (basado en Perov et al., 2020).

O bien otra tecnología más moderna, que es la del ultrafalso o *deepfake* (Figura 6), en la cual el sistema aprende a dibujar la cara de la persona base mediante inteligencia artificial GAN (redes neuronales adversariales). El método o tecnología del ultrafalso consiste en un entrenamiento especial, donde se generan caras en la parte generativa, y se decide si la cara es real o es falsa (*fake*) en la parte adversarial. Mediante sistemas de codificación y decodificación el sistema va aprendiendo técnicas de dibujado de aquellas caras que mejor se ajustan. El sistema aprende también a trabajar con colores y a "pintar las caras". Para cada cara base tenemos que entrenar el modelo, por lo que el número de iteraciones puede ser muy elevado y requerir horas o incluso días de trabajo.

Figura 6. Método o tecnología ultrafalsa o deepfake para el intercambio de caras



Fuente(s): Euskal Encouter, 2020, 12m42s (basado en Perov et al., 2020).

En cualquier caso, a la hora de elegir entre los dos métodos (el tradicional de triangulación, propio del *face swap*; o el del ultrafalso o *deepfake*) debemos tener en cuenta los siguientes aspectos (Tabla 3):

Tabla 3. Comparativa entre los métodos de triangulación y del ultrafalso en el intercambio de caras

Triangularización propio del *face swap*	**Ultrafalso propio del d*eepfake***
Fallos de renderizado (o de representación gráfica), totalmente dependiente de los *landmarks* (o puntos de referencia).	Resultados impresionantes.
Se puede hacer en tiempo real.	Solo se puede usar en vídeos e imágenes.
No necesita entrenamiento.	Necesita un entrenamiento previo.

Fuente(s): Adaptación de Euskal Encouter (2020), 25m15s (basado en Perov et al., 2020).

En el caso de la clonación o intercambio de caras tradicional (o *face swap*), para poder obtener unos resultados finales "más naturales", se acostumbra a aplicar un suavizado en los píxeles. Este suavizado se realiza mediante una suma ponderada de las dos matrices de la imagen, aplicándole la transparencia. En el caso de la clonación o intercambio de caras mediante la tecnología del ultrafalso (o *deepfake face*), como veremos en el ejemplo que analizaremos en el apartado de "Resultados", hoy en día, pese a usar la inteligencia artificial para su creación, el proceso sigue siendo requiriendo de algunos "ajustes manuales" de pre o posproducción si queremos unos resultados de calidad óptimos. Cabe recordar a este respecto que este es uno de los motivos por los cuales optamos por analizar ejemplos profesionales realizados con esta tecnología en vez de crear nosotros mismos nuestros propios ejemplos.

3.2.3. La traducción cara a cara

En los dos subapartados anteriores hemos analizado como utilizar la tecnología del ultrafalso o *deepfake* para la clonación o intercambio de audios (voces) y de imágenes (caras); por separado, o integradas en lo vídeos. Analizaremos, a continuación, si es posible conjugar ambas tecnologías del ultrafalso para traducir o doblar vídeos ultrafalsos. De ser el caso, lo que *a priori* parece más procedente es que ambos modelos (audio-voces, imágenes-caras) del ultrafalso sean, cada uno de ellos por separado, un componente diferenciado del modelo ultrafalso integral (vídeo con voces y caras; con audio e imágenes en movimiento o fotogramas) final; y que a estos se les añada un tercer módulo relacionado con la traducción/doblaje propiamente dicho, donde la sincronización labial sea uno de los aspectos más importantes. Siempre y cuando, obviamente, queramos que la calidad de los resultados en los vídeos sea la mejor posible. Además, en aras a su posible implementación en el ámbito de la traducción/doblaje de vídeos educativos en el ámbito de la enseñanza superior (universitaria) virtual, en general, y en el ámbito el Campus Remoto e la UVigo, en particular; abogaremos por un sistema integral resultante que sea efectivo (automático o lo más automático posible), barato (gratuito, a ser posible) y que produzca soluciones de calidad (o con la mayor calidad posible).

Hoy en día Internet está lleno de videos de caras que hablan: YouTube (con 300 horas subidas por día), conferencias en línea, videoconferencias (como las creadas en el Campus Remoto por los docentes de la UVigo), películas, programas de televisión, etc. A este respecto, K. R. Prajwal, el investigador principal del sistema integrado para la traducción de ultrafalsos ajustados a la sincronía labial denominado traducción cara a cara o *face-to face translation* (Prajwal et al., 2019b) que analizaremos en el apartado de "Resultados", consideraba que los sistemas de traducción existentes hasta ese momento (antes de su propuesta de traductor de caras) solo podían generar una salida de voz traducida o subtítulos textuales para dicho contenido de video; pero no eran capaces de manejan el componente visual. Como resultado de ello, el discurso traducido en dichos sistemas (anteriores a su propuesta), cuando se superponía al video, los movimientos de los labios no estaban sincronizados con el audio.

Una vez explicadas y analizadas las tecnologías del ultrafalso de voz (audio) y de cara (imágenes), ya sea por separado o de manera conjunta en los vídeos, y, en el caso de los vídeos, explicada y analizada la tecnología de la traducción cara a cara que nos permitirá la traducción/doblaje automatizado y de calidad de los vídeos originales; en el siguiente apartado de "Resultados" analizaremos algunos ejemplos prácticos ya existentes relacionados con estas tres tecnologías (un ejemplo por tecnología).

4. Resultados

En este apartado expondremos los resultados obtenidos a partir de la implementación de las tres tecnologías expuestas y analizadas en el apartado anterior.

Para ello, en primer lugar, profundizaremos un poco más en el funcionamiento y utilidad de cada una de las tecnologías analizadas; para, a continuación, mostrar un ejemplo de estas. Como ya hemos comentado anteriormente, los ejemplos que mostraremos son ejemplos profesionales de ámbitos que no se corresponden con nuestro ámbito de estudio que es el de los vídeos educativos en la educación superior (universitaria) virtual. En cualquier caso, una vez mostrados los resultados de la aplicación de dichas tecnologías en estos ámbitos, trataremos de justificar su potencial uso en nuestro ámbito particular de estudio.

En lo que concierne la tecnología del ultrafalso de voz, en general, "en la actualidad, la demanda por parte del mercado de la clonación de voz es cada vez mayor, debido a sus interesantes y variadas aplicaciones (Vicomtech, 2022a)", entre las que se encuentran los asistentes conversacionales (tipo Alexa), los altavoces inteligentes (tipo Takara Tomi, comentado anteriormente), el doblaje (objetivo central de este trabajo), los personajes digitales, los juegos, los audiolibros (objetivo central del proyecto de investigación asociado a este trabajo), los sistemas de navegación o el *branding* de voz.

El *branding* "es un anglicismo empleado en mercadotecnia que hace referencia al proceso de hacer y construir una marca ("Branding", 2022)". En la actualidad resulta fundamental trabajar en esta área para dotar a las marcas de una identidad sonora. "A día de hoy, pocos proyectos de *branding* tienen sus propios sonidos, voz (hablada) y música propia que les identifiquen en contextos donde no hay visuales para apoyar el mensaje (Vilardi, 2019)".

La tecnología ultrafalsa de voz ayudará a automatizar y personalizar muchas tareas realizadas en este tipo de aplicaciones y dominios.

En el ámbito del doblaje y de la animación digital, en particular, "la clonación de la voz permite crear voces identificables para personajes digitales únicos o avatares en sistemas de interacción entre persona y ordenador, e incluso para la producción de contenidos multimedia y audiovisuales (Vicomtech, 2022a)".

En lo que concierne la integración de la tecnología del ultrafalso de voz y la tecnología del ultrafalso de cara en un mismo producto, el vídeo digital; para poder realizar un vídeo ultrafalso, *a priori* (al ser un vídeo una sucesión de imágenes estáticas o fotogramas puede partirse también de una o varias imágenes estática; o bien, si se trata solo de cambiar el mensaje de voz utilizando el mismo vídeo de caras, solo se precisaría del vídeo original y del nuevo audio –o incluso solo del texto, si es que vamos a realizar conversión de voz-a-texto, traducción texto-a-texto, y, conversión texto-a-voz–) es necesario disponer de dos vídeos, uno para la persona que se quiere suplantar (que, en nuestro caso, serían los vídeos originales generados por el personal docente en el Campus Remoto; poseditados para mejorar su calidad y adaptarlos a longitudes "asequibles") y el otro para la persona suplantada (que, en nuestro caso, serían únicamente otros audios -o el texto a convertir en audio- del propio personal docente).

El proceso automático de generación de un ultrafalso consiste en utilizar una red generativa adversarial (RGA) basada en inteligencia artificial, entrenada con millones de imágenes de caras. En dicha red neuronal se realiza un entrenamiento de las caras de los 2 vídeos (o de la cara del primer vídeo con los audios en el mismo u otros idiomas –audios ya generados *a priori* o que se generarían *a posteriori* a partir de textos en otras lenguas– del segundo vídeo), realizando ajustes por cada iteración.

En la red existe la parte generativa y la parte adversarial. La parte generativa es la encargada de crear el ultrafalso propiamente dicho, mientras que la parte adversarial es la encargada de decidir si esa cara es real o es falsa (es decir, si es un *fake*). Cuando el sistema adversarial es engañado por el generativo, la cara se va suplantando en el video fotograma a fotograma.

El coste temporal de la creación de ultrafalsos relacionados con la traducción (o doblaje) de vídeos educativos es todavía demasiado elevado para las necesidades requeridas (ya no digamos para poder realizarlo en tiempo real); ya que, en general, los docentes universitarios disponen de poco tiempo para su realización y, además, el número de vídeos educativos generados en las grabaciones de las plataformas docentes virtuales es muy elevado (mucho más, si además los asociamos a la "regla 2-6-10" de optimización de los vídeos educativos, que establece que el tamaño óptimo de un vídeo educativo es de 6 minutos como máximo). Sin embargo, si observamos los resultados obtenidos en los ejemplos que mostraremos a continuación, donde el factor tiempo no es un condicionante, los resultados son muy esperanzadores.

4.1. Ejemplo de ultrafalso de voz

El ultrafalso de voz que mostramos a continuación (Cámara, 2022; Del Barco, 2020; Entrepelis, 2022; Spotify, 2020a, 12m52s; Spotify, 2020b, 30m30s) está relacionado con la figura histórica, ya fallecida, del dictador español Francisco Franco ("Francisco Franco", 2022). La serie de podcasts de Spotify (2020a, 12m52s; 2020b, 30m30s) titulada XRey, relacionada con la serie documental *Los borbones: una familia real* (Cámara, 2022; Entrepelis, 2022), requirió del uso de ultrafalsos de voz en dos de sus capítulos. En el cuarto capítulo "se puede escuchar al dictador leyendo la carta escrita a Juan de Borbón, padre del rey emérito, comunicándole la decisión de proponer a Juan Carlos como su sucesor en la jefatura del Estado (Del Barco, 2020)". Este fragmento de 27 segundos fue recreado "de maneara artificial, generando un *deepfake* o simulación que se asemeja en extremo a la voz de Franco. Con ello se consigue un recurso que añade riqueza a la historia y acerca al oyente a lo narrado (Del Barco, 2020)". El ultrafalso de voz está disponible en el podcast de Spotify (2020b, 28m39s-29m06s); mientras que, en el podcast de Spotify (2020a, 12m52s) se explica cómo se realizó dicho ultrafalso de voz con ayuda de la inteligencia artificial (Figura 7). La empresa encargada de la realización de dicho ultrafalso de voz fue Vicomtech (2022a, 2022b, 2022c).

Figura 7. Ejemplo de ultrafalso de voz (con la recreación de la voz de Francisco Franco)

Fuente(s): Spotify, 2020a, 12m52s; Spotify, 2020b, 30m30s.

Según se escucha en el *making off* de Spotify (2020a, 12m52s) del ultrafalso de voz de Spotify (2020b, 30m30s), donde se explica la creación de dicho ultrafalso de voz; como recurrir a un actor de doblaje o a un imitador de la voz de Franco no tenía nada de novedoso, los productores de la serie decidieron optar por la fórmula del ultrafalso de voz. Tras una intensa búsqueda contactaron con Vicomtech (2022a), una empresa especializada en inteligencia artificial afincada en Donostia / San Sebastián (España), para llevar a cabo dicho trabajo. Para dicha empresa, "la idea principal era entrenar, construir, un sistema, que tomara como entrada un texto; y, en su salida, generara un audio, locutando dicho texto (Spotify, 2020b, 06m11s-06m23s)". Además, "ese audio tenía que tener las características vocales, de entonación, matices, etc. de Franco (Spotify, 2020b, 06m25s-06m31s)". Para ello, usaron técnicas de inteligencia artificial para construir ese tipo de sistemas; que, "para generarse, para entrenarse, necesitan muestras similares a las que después van a generar. Si yo voy a generar voces de Franco, lo que necesito para entrenar el sistema son voces de Franco (Spotify, 2020b, 06m46s-06m55s)".

La mayor dificultad estuvo, según ellos, en el proceso de entrenamiento del sistema. Hacerlo con personajes actuales o de los que existen registros en alta calidad no era *a priori* un problema, pero cuando hay que recurrir a grabaciones de varias décadas de antigüedad, como es el caso, todo se complica mucho más. En un primer momento trabajaron con "no menos de 20 horas de audios (Spotify, 2020b, 10m46s-10m50s)". Sin embargo, la dificultad de encontrar audios en buen estado, limpios de ruidos y en el estilo narrativo que se buscaba, hizo que finalmente el modelo tuviera que generarse con "tan solo 6 horas (Spotify, 2020b, 10m55s-11m06s)"; compuestas, principalmente, por discursos navideños del dictador, "con voces del [año] 55 al 69 (Spotify, 2020b, 10m20s-11m25s)".

Los resultados obtenidos consiguen alcanzar la calidad deseada; e incluso, en el *making off* (Spotify, 2020a, 13m47s-13m56s), como epílogo el podcast, la inteligencia artificial pone a Francisco Franco a cantar la canción de *La Macarena* (YouMoretv-Espectáculo, 2014).

4.2. Ejemplo de ultrafalso integral (de voz y de cara)

El ultrafalso integral (de voz y de cara) que mostramos a continuación (CruzcampoTV, 2021a, 2021b; Lacort & Tones, 2021; Panorama Audiovisual, 2021) está relacionado con la figura histórica, ya fallecida, de la cantante, bailaora y actriz española María Dolores Flores Ruiz, más conocida como Lola Flores ("Lola Flores", 2022). Se trata de un vídeo de 1 minuto de duración de un anuncio promocional de la empresa cervecera Cruzcampo (Heineken España). En este caso, disponemos del ultrafalso del anuncio y de un vídeo creado *a posteriori* donde se explica cómo se creó dicho vídeo (Figura 8). El ultrafalso fue creado por la empresa Metropolitana (2022), radicada en Barcelona y Madrid (España); que es una empresa de post producción VFX (siglas en inglés de efectos visuales), publicidad, cine y televisión.

Figura 8. Ejemplo de ultrafalso integral, de voz y de cara ("Con Mucho Acento", de Lola Flores)

Fuente(s): CruzcampoTV, 2021a, 2021b; Panorama Audiovisual, 2021.

Según Panorama Audiovisual "gracias a la tecnología basada en inteligencia artificial, Ogilvy [la agencia publicitaria creadora del proyecto] y Metropolitana [el estudio de efectos especiales y postproducción que lo llevó a cabo] conciben una campaña 'protagonizada' por un *deepfake* [integral, de voz y de caras] de Lola Flores 25 años después de su desaparición (Panorama Audiovisual, 2021)".

A este respecto, en una conversación de Lacort & Tones (2021) con Pedro Moreno y Javier Senovilla (de la empresa Ogilvy, ambos directores creativos del anuncio), Ramón Arteman (director y cofundador de Metropolitana) y Nico Roig (miembro del departamento de 3D de Metropolitana); todos ellos proporcionaron múltiples detalles del proceso seguido para la realización de dicho ultrafalso integral (de voz y de cara). En cuando a las cuestiones "de voz" podemos destacar:

> Lolita imitó la voz de su madre en un proceso que también pasó por una profunda postproducción. "Luego íbamos experimentando con la grabación", afirma Moreno, "haciendo pruebas con limitadores para modificar el tono, y mediante un software especial fuimos modulando el timbre para acercarlo al original" tiene el timbre y el acento inconfundible de la matriarca de los Flores. (Lacort & Tones, 2021)

Y, en cuando a las cuestiones "de cara" podemos destacar:

> Para empezar, Metropolitana tuvo que recopilar imágenes que sirvieran de base: "La idea era tener la mayor cantidad posible de rangos de ángulos de la cara y expresiones. Todas las horas que teníamos de metraje se reducían a unas 5.000 imágenes aproximadamente con las que alimentamos la inteligencia artificial", cuenta Ramón Arteman, director y cofundador de Metropolitana. [...] Nos presentaron unas cien fotos y nos preguntaron cuál nos gustaba. Cogimos una imagen de Lola que fue como un tótem", explica Ramón.
>
> El siguiente paso fue grabar a una actriz creando un modelo 3D, "un molde sobre el que se iría tejiendo un entramado quirúrgico sobre el que se recrearía el rostro de Lola", nos comenta Javier Senovilla, uno de los directores creativos del spot. La actriz sería la base de todos los gestos sobre los que luego irían las imágenes de la Lola Flores real, "que luego se perfilaría retocando digitalmente expresiones de la cantante y detalles de la piel, como la parte de la unión del cabello con la frente, que hubo que retocar por separado para salvar las diferencias entre el físico de la modelo y el de Lola Flores", afirma Roig. En este proceso se usaron dos programas específicos para crear deepfakes. Por un lado, FaceSwap [2022], un software de código abierto basado en TensorFlow, Keras y Python. Tiene una apariencia muy visual que facilita el trabajo, al menos en una primera fase. Para la segunda fase se requería algo que permitiese profundizar mucho más en los

detalles, de forma más artesanal, y usando código. Ese segundo software fue DeepFaceLab [2022], utilizado para esos largos retoques finales.

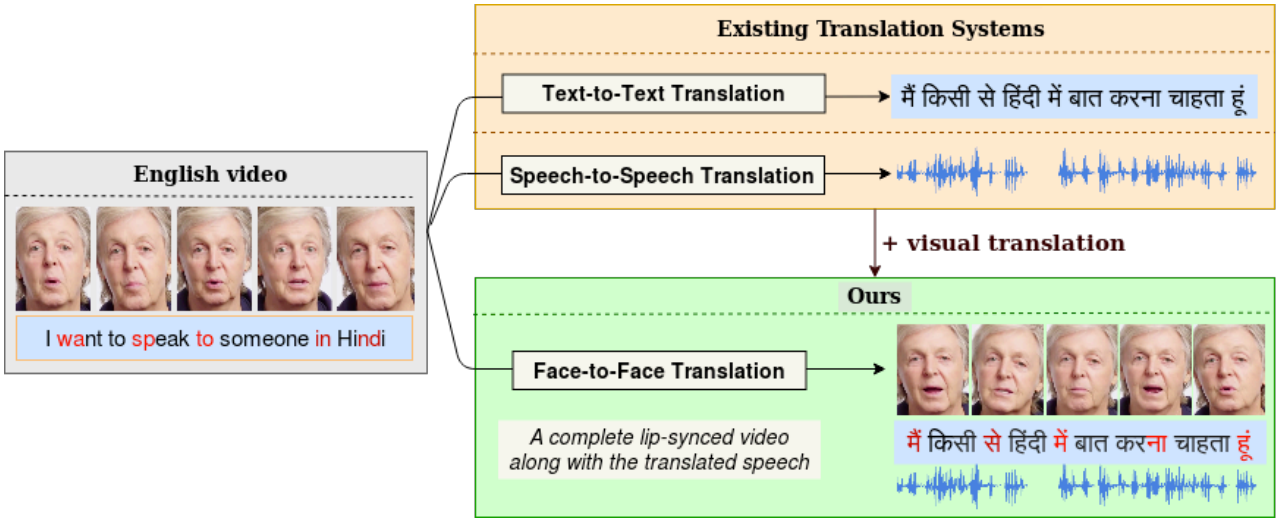
Nico Roig, del departamento 3D de Metropolitana y uno de los encargados de devolver a Lola Flores a la vida en cierta forma, explica que "DeepFaceLab nos daba mejor resultado a la hora de controlar todos los parámetros, que es el quid de la cuestión", y que "el retoque final, no solo del deepfake sino de su integración final, composición, etc.; nos llevó casi tanto tiempo como el propio deepfake. El último 10% supuso casi tanto tiempo como todo lo anterior". (Lacort & Tones, 2021)

4.3. Ejemplo de traducción cara a cara de ultrafalsos

Tomando en consideración los avances en los sistemas de traducción automática, Prajwal et al. (2019a, 2019b) y Rudrabha Mukhopadhyay (2019) proponen un enfoque novedoso en lo que denominan traducción cara a cara o *face to face translation*, relacionada con la traducción (o doblaje) de ultrafalsos. Argumentan que, "a medida que la comunicación digital actual se vuelve cada vez más visual, existe la necesidad de sistemas que puedan traducir automáticamente un video de una persona que habla en el idioma A a un idioma de destino B, con una sincronización labial realista (García Nieto, 2020)".

Para ello, en primer lugar, proponen la creación de un sistema de traducción de voz a voz funcional que reúne varios módulos de voz y lenguaje ya existentes. Luego, avanzan hacia la "traducción cara a cara" mediante la incorporación de un nuevo módulo visual, que denominaron LipGAN, para generar caras parlantes realistas a partir del audio traducido.

Figura 9. Esquema básico del sistema de traducción cara a cara

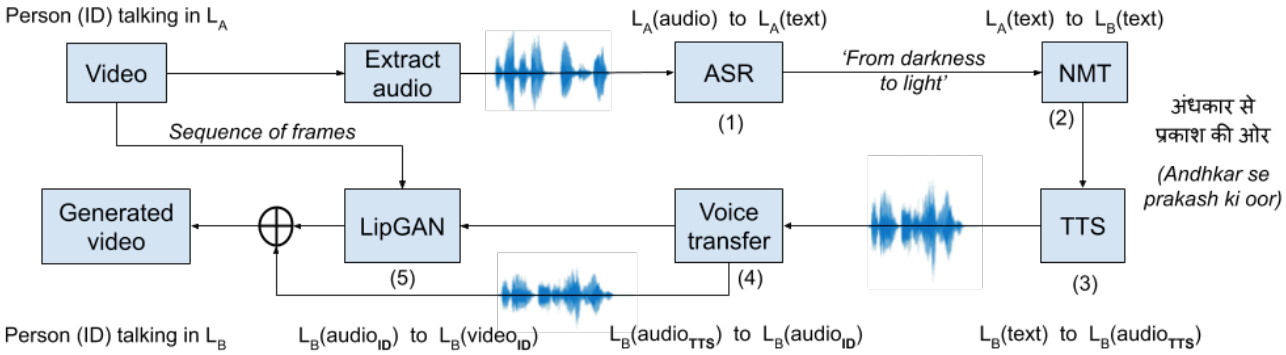


Fuente(s): Prajwal et al., 2019a.

En la Figura 9 de la página anterior ilustramos una posible aplicación en el mundo real de un sistema de este tipo, propuesta por Prajwal et al. (2019a, 2019b), donde dos personas pueden entablar una conversación natural en sus respectivos idiomas. Se trata de un modelo integral de traducción de voces y de caras, donde un determinado hablante habla en un idioma I1 de una lengua L1 (hindi, en este caso) en un vídeo inicial; y, el sistema en cuestión, totalmente automatizado, genera un video del hablante hablando en un idioma I2 de una lengua L2 (inglés, en este caso).

En el modelo de Prajwal et al. (2019a, 2019b) ya implementado, observamos que dicho sistema se divide, a su vez, en dos subsistemas: un primer subsistema de traducción de voz a voz, y, un segundo subsistema de síntesis de caras del sujeto que incluye también un sistema de síntesis de labios.

Figura 10. Esquema del subsistema de traducción de voz a voz del modelo de traducción cara a cara



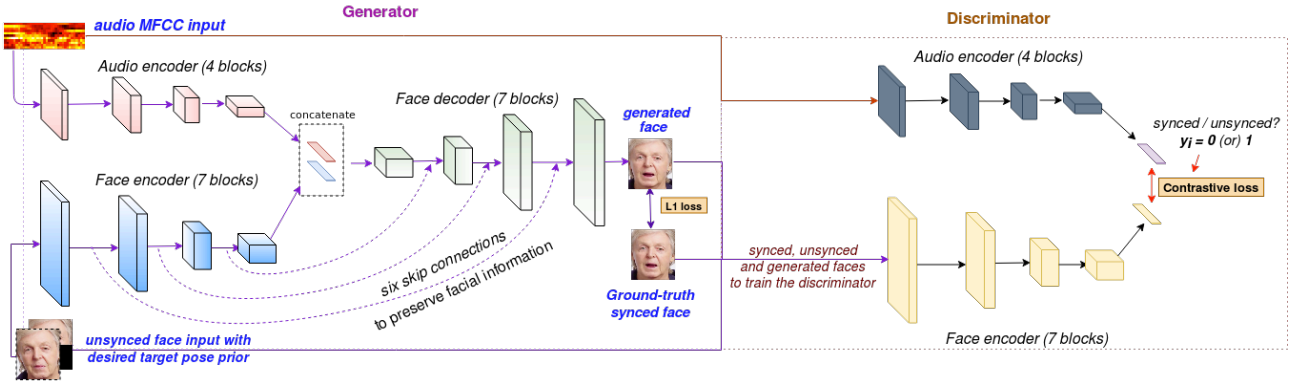
Fuente(s): Prajwal et al., 2019a, 2019b.

El primer subsistema de traducción de voz a voz (Figura 10) combina tecnologías como ASR (reconocimiento automático del habla, por sus siglas en inglés de *Automatic Speech Recognition*), NMT (traducción automática neuronal, de *Neural Machine Translation*) y TTS (conversión de texto a voz, de *Text-to-Speech*):

> Para, primero, entender lo que dice el audio original; segundo, traducirlo a otro idioma y tercero, "hablarlo" en el idioma objetivo usando la voz del sujeto. Para traducir el inglés se ha usado DeepSpeech [8] y para entrenar el sistema NMT en los idiomas indios se ha usado el código base de Facebook AI Research [9]. (García Nieto, 2020)

El segundo subsistema de síntesis de caras y de labios (Figura 11) incorpora un nuevo modelo creado *exprofeso*, denominado LipGAN.

Figura 11. Esquema del subsistema de síntesis de caras y de labios del modelo de traducción cara a cara



Fuente(s): Prajwal et al., 2019a, 2019b.

Al respecto de LipGAN, se trata de:

> Un modelo que puede generar rostros parlantes de cualquier persona a partir de cualquier fragmento de audio. El modelo consta de dos codificadores: un codificador facial y un codificador de voz. El primero se usa para "codificar información sobre la identidad del sujeto", tal y como explican los autores, mientras que el segundo "toma fragmentos de voz muy pequeños (350 ms de audio por fragmento) y los usa para codificar la información de audio". Finalmente, el output de estos dos codificadores se introduce en un decodificador que genera una imagen facial con los labios sincronizados con el fragmento de audio. (García Nieto, 2020)

El módulo LipGAN fue sometido a dos evaluaciones. En la evaluación cuantitativa se usaron diferentes métricas estándar y se enfrentó LipGAN a otros modelos similares, consiguiendo LipGAN la mejor puntuación en todos los casos. En la evaluación cualitativa humana, 20 personas evaluaron el realismo y la sincronización labial de 150 ejemplos en una escala de uno a cinco; y, de nuevo, LipGAN consiguió la puntuación más alta (Chung & Zisserman, 2016; García Nieto, 2020; Prajwal et al., 2019a).

El proceso de traducción cara a cara fue sometido también a múltiples evaluaciones humanas y estas demostraron que puede mejorar significativamente la experiencia general del usuario para consumir e interactuar con contenido multimodal en varios idiomas. Según García Nieto (2020) "de acuerdo a los investigadores, y si bien

8 Disponible, como código abierto, en https://github.com/mozilla/DeepSpeech
9 Disponible, como código abierto, en https://github.com/facebookresearch/fairseq

el software no es lo suficientemente rápido para hacer esta traducción en tiempo real, LipGAN tiene potencial para el doblaje de películas, de vídeos educativos y para entrevistas y noticias en la televisión (García Nieto, 2020)".

Para entender mejor cómo funciona todo el proceso puede consultarse el vídeo de demostración creado *exprofeso* por los propios investigadores y disponible en Rudrabha Mukhopadhyay (2019). En la marca de tiempo 06m38s-06m46s del vídeo en cuestión, por ejemplo, puede visualizarse un videoclip de la difunta princesa Diana de Gales en una entrevista de 1995 con el periodista Martin Bashir, pronunciando, en inglés la famosa frase "I'd like to be a queen in people's hearts but I don't see myself being queen of this country". Un momento después, en la marca de tiempo 06m46s-06m54s, aparece otra vez Diana de Gales pronunciando la misma cita en hindi, con los labios moviéndose, como si ella realmente hablara ese idioma (Figura 12).

Figura 12. Ejemplo de traducción cara a cara de ultrafalsos (relacionado con Lay Di)

Fuente(s): Adaptado de Rudrabha Mukhopadhyay, 2019, 06m38s-06m46s.

5. Discusión

A partir de lo expuesto anteriormente, de los ejemplos mostrados y de la literatura relacionada con las tecnologías del ultrafalso (o *deepfake*) y de la traducción cara a cara (o *face-to-face-translation*) podemos concluir que en la actualidad existen los conocimientos, las habilidades y la tecnología informática necesarias para poder crear desde cero (a partir de textos) vídeos en diferentes idiomas; y/o editar vídeos ya creados (reaprovechándolos) en un idioma determinado y traducirlos/doblarlos en otros idiomas diferentes a los utilizados en el momento de su creación; todo ello con una calidad aceptable. Aunque hemos constatado también que (todavía) no posible realizar este proceso en tiempo real.

Si bien los ejemplos prácticos de implementación de dichas tecnologías mostrados en este trabajo no están directamente relacionados con nuestro ámbito de estudio (por las razones que ya hemos expuesto en diversas ocasiones) de la generación y traducción/doblaje de vídeos educativos en la enseñanza superior (universitaria) virtual, en general, y en el ámbito del Campus Remoto de la UVigo, en particular; creemos que dichas tecnologías poseen el potencial y son extrapolables a nuestro ámbito específico de análisis. De entre toda la literatura que apoya nuestras conclusiones destacamos el artículo "El lado bueno del 'deepfake': el doblaje perfecto para series y películas" que José María López Lafuente (2020), experto en nuevas tecnologías, escribió acerca del doblaje de series y películas:

> Obviamente, cuando hablamos de tecnología de manipulación de imágenes, es inevitable que surja el concepto de inteligencia artificial. Gracias a este campo tecnológico hemos sido capaces de avanzar rápidamente en el procesamiento de datos, imágenes y, ahora, de vídeos. Todo ello con ayuda de computadoras, software y los algoritmos adecuados. Sin presupuestos desorbitados como pudiera ocurrir en el pasado […] Esta rama de la inteligencia artificial [*machine learning*] hace posible que el propio software sea el encargado de discriminar los resultados poco satisfactorios. Además, hay que tener en cuenta la complejidad de sincronizar audio y vídeo, algo imposible hasta hace poco.

> Aunque los resultados son satisfactorios, todavía queda trabajo por hacer. En cualquier caso, el software ahí está, a disponibilidad de quien quiera usarlo. Y como indican sus responsables, su utilidad es variada, tanto para el doblaje de películas como para crear material educativo en varios idiomas o para adaptar contenido televisivo como noticias o entrevistas. (López Lafuente, 2020)

En cualquier caso, en un ultrafalso o *deepfake* integral (de voz y cara) de vídeos educativos hay que manipular y sincronizar audio y vídeo; haciendo hincapié en la sincronización labial si queremos obtener vídeos naturales. Es por ello por lo que nos hemos interesado en modelos integrales como es el caso del modelo de traducción cara a cara de Prajwal et. al. (2019b); que tiene en cuenta todos estos aspectos.

6. Conclusiones

En este trabajo hemos podido llevar a cabo todos los objetivos que nos habíamos planteado al inicio de este. En primer lugar, hemos podido constatar, de manera teórica y de manera práctica (mediante ejemplos de su implementación), que la tecnología del ultrafalso o *deepfake* y la tecnología de la traducción cara a cara o *face-to-face translation* pueden utilizarse en el ámbito de la traducción/doblaje de vídeos (de caras), en general.

Aunque los ejemplos prácticos que hemos tomado como muestra no pertenezcan al ámbito específico objeto de estudio de la enseñanza superior (universitaria) virtual; sin embargo, creemos que *a priori*, nada parece indicar que no sea posible hacerlo cuando la tecnología evolucione y mejore y el factor tiempo no sea tan condicionante como lo es ahora.

Por otra parte, hemos podido constatar también que, al ser tanto los programas y las librerías asociados a dichas tecnologías como el propio Campus Remoto de la UVigo de código abierto, tampoco *a priori* nada parece indicar que no se puedan integrar dichas tecnologías en dicha plataforma de teledocencia. De ser el caso, el personal docente de la UVigo, que en este momento ya puede grabar sus clases docentes de manera automatizada y generar sus propios vídeos educativos, podrá también traducir/doblar dichos videos a otros idiomas diferentes a los de los vídeos originales de manera automatizada y a coste cero.

La tecnología el ultrafalso permitirá generar (clonar o intercambiar) nuevas voces y caras (en nuestro caso de los mismos docentes) a partir de otros vídeos; mientras que la tecnología de la traducción cara a cara permitirá asegurar que el proceso de ajuste labial (y/o facial) entre dichas nuevas caras y voces obtenidas al final de todo el proceso de traducción/doblaje sea óptimo, de manera que se consiga la mayor naturalidad posible en el discurso resultante.

En términos de aplicabilidad y relacionado con el ámbito de la traducción/doblaje de vídeos educativos podemos concluir, por tanto, que, hoy por hoy, es ya posible traducir/doblar vídeos educativos (con un solo personaje) en un tiempo relativamente corto y con unos resultados óptimos. Si bien el software a emplear no es (todavía) lo suficientemente rápido para hacer esta traducción/doblaje en tiempo real (lo cual, como ya hemos dicho, *a priori* no es impedimento alguno para nuestro propósito de traducción de vídeos educativos), los investigadores que han desarrollado esta tecnología y programas asociados consideran que tiene potencial para, no solo la traducción/doblaje de vídeos educativos de manera automática y en tiempo real, sino también para el doblaje de películas y el doblaje de entrevistas o de noticias en televisión o en otros medios (por ejemplo, los OTT, como Twitch).

Además, como se desprende de nuestro trabajo, la tecnología y los programas informáticos necesarios para poder llevar a cabo está labor son todos ellos gratuitos y, la mayoría, de código abierto; lo cual hace prever que el avance en la mejora de la tecnología y del proceso global de traducción/doblaje será mucho mayor y en un lapso menor. En Prajwal et al. (2019a) puede accederse a toda la información básica relacionada con dicha tecnología: el artículo de investigación teórico que sirvió de base a dicha tecnología, los programas informáticos (librerías) de código abierto a usar; y, finalmente, un muy interesante e ilustrativo vídeo demostrativo acerca los principios básicos de dicha tecnología y de su uso práctico (con ejemplos varios).

Por nuestra parte, como docentes de la UVigo y usuarios de la plataforma telemática (y de grabación y edición de vídeo) del Campus Remoto de la propia UVigo, hemos contactado ya con los Servicios Informáticos de la UVigo que se ocupan de la implementación y del mantenimiento de dicho Campus Remoto para comentarles la posibilidad de incorporar (ahora o en un futuro no muy lejano) este servicio de "traducción/doblaje cara a cara" de los vídeos educativos generados mediante las grabaciones (automáticas) de las clases docentes, conferencias, etc. llevadas a cabo en dicha plataforma.

7. Agradecimientos

El presente texto nace en el marco del proyecto EACEA-32-2019-COOP1 "G-Book 2: European Teens as Readers and Creators in Gender-positive Narratives (G-Book 2)", en el cual participa la UVigo, financiado por la Unión Europea (convocatoria Creative Europe Programme 2014-2020).

Referencias

AWS Events. (2022, 23 de junio). *Amazon re:MARS 2022 - Day 2 – Keynote* [vídeo]. YouTube. https://cutt.ly/zLk78pc

Ballantine, M. (2019, 3 de julio). *Are Deepfakes Invading The Office?* Forbes. https://cutt.ly/HK0GY8K

Benítez, S. (2022, 5 de julio). *Disney: 'Blancanieves y los siete enanitos', de David Hand*. Espinof. https://cutt.ly/bL5Sdha

Bmotion (2022). ¿Qué es un fotograma? *Glosario audiovisual*. https://cutt.ly/ILk5MQR

Branding. (15 de marzo de 2022). En *Wikipedia.* https://cutt.ly/WL2RMVV

Brown, D. (2019, 13 de mayo). *Wait, is that video real? The race against deepfakes and dangers of manipulated recordings*. USA Today Tech. https://cutt.ly/kK0FQmE

Cámara, N. (2022, 27 de junio*). 'Los Borbones: una familia real': dónde ver la serie documental completa*. Diezminutos. https://cutt.ly/xK60O5x

Chung, J. S. & Zisserman, A. (2016). Lip reading in the wild. En *Asian Conference on Computer Vision*. Springer, 87–103. https://cutt.ly/4LjVQgs

Cole, S. (2017, 11 de diciembre). *AI-Assisted Fake Porn Is Here and We're All Fucked*. Motherboard, Vice. https://cutt.ly/HK8LVye

Costa, E.- (2022, 26 de abril). *Snapchat: número de usuarios sigue creciendo al igual que sus ingresos en anuncios*. Noticias MetroCr. https://cutt.ly/TK5zdrk

CruzcampoTV. (2021a, 21 de enero). *Cruzcampo. Así se hizo #ConMuchoAcento* [vídeo]. YouTube. https://cutt.ly/7Lk5eJl

CruzcampoTV. (2021b, 21 de enero). *Cruzcampo. Con mucho acento* [vídeo]. YouTube. https://cutt.ly/CLk5kbk

DeepFaceLab. (2022). *DeepFaceLab* [software]. https://cutt.ly/oLk6NFY

Deepfake. (21 de junio de 2022). En *Wikipedia*. https://cutt.ly/8L2Tyku

Del Barco, L. (2020, 3 de junio). *Spotify "resucita" a Franco creando un 'deepfake' de la voz del dictador para su nuevo podcast*. Hipertextual. https://cutt.ly/ZK62sTV

Dickson, B. (2018, 7 de junio). *When AI Blurs the Line Between Reality and Fiction*. PCMag. https://cutt.ly/OK0O1rc

Ebert, R. (2009, 12 de agosto). *Finding my own voice*. RogerEbert. https://cutt.ly/AK0Appi

Entrepelis (2022, 31 de mayo). *Los Borbones: una familia real. 1x1 online y en castellano* [vídeo]. Netu TV. https://cutt.ly/1Lqybnu

Euskal Encouter. (2020, 5 de diciembre). *#AE07. DeepFakes y el Face Swap. ¿Cómo funcionan los algoritmos de suplantación?* [vídeo]. YouTube. https://cutt.ly/BLk5n6S

FaceSwap. (2022). *FaceSwap* [software]. https://faceswap.dev/

FakeYou. (s.f.). *FakeYou* [software]. https://fakeyou.com/

Figueira, A., & Oliveira, L. (2017). The current state of fake news: challenges and opportunities. *Procedia Computer Science, 121*: 817–825. https://doi.org/10.1016/j.procs.2017.11.106

Fotogramas por segundo. (18 de marzo de 2022). En *Wikipedia.* https://cutt.ly/dL2Yj48

Francisco Franco. (20 de julio de 2022). *Wikipedia*. https://cutt.ly/VL2Y8Tn

García Nieto, J. (2020, 19 de febrero). *Este software es capaz de generar vídeos deepfake traducidos a diferentes idiomas con los labios sincronizados*. Xataka. https://cutt.ly/BLrg9aa

Harding, S. (2022, 23 de junio). *Amazon uses kid's dead grandma in morbid demo of Alexa audio deepfake*. ArsTechnica. https://cutt.ly/aK2HKZR

Ion, F. (2022, 22 de junio). *Amazon's Latest Trick: Pipe In a Dead Person's Voice Through Alexa's Speakers*. Gizmodo. https://cutt.ly/4K2Llti

KULTT [@kultt_fr]. (22 de abril de 2016). *Dans mon quartier, affichage @NetflixFR pour réaliser des Face Swap avec ses héros #NetflixSwap* [tuit]. Twitter. https://cutt.ly/hK2I5Ma

Kuper, S. (2018, 18 de octubre). *The age of scepticism: from distrust to 'deepfake'*. Financial Times. https://cutt.ly/QK0HrYk

La Región (7 de noviembre de 2021). O Campus Remoto da UVigo suma 2,5 millóns de conexión. *La Región*. https://cutt.ly/YLrbYVP

Lacort, J. & Tones, J. (2021, 21 de enero). *Hablan los creadores del 'deepfake' de Lola Flores: "Pese a usar Inteligencia Artificial fue un proceso bastante artesanal".* Xataka. https://cutt.ly/YK0A3I5

Lola Flores. (12 de julio de 2022). En *Wikipedia*. https://cutt.ly/rL2UGUW

López Ardao, C. (2017, 16 de octubre). *¿Cuál debería ser la longitud del vídeo educativo?* The Flipped Classroom. https://cutt.ly/sLlwvdQ

López Lafuente, J. M. (2020, 10 de octubre). *El lado bueno del 'deepfake': el doblaje perfecto para series y películas.* Hipertextual. https://cutt.ly/zLwNLjj

Marr, B. (2019, 22 de julio). *The Best (And Scariest) Examples Of AI-Enabled Deepfakes*. Forbes. https://cutt.ly/vK0OcsP

Metropolitana (2022). *Metropolitana. Post producción VFX, publicidad, cine y televisión.* Metropolitana. https://cutt.ly/vLlwWNq

Miquel-Vergés, J. (2022a) Enseñanza-aprendizaje virtual en el ámbito universitario en tiempos de pandemia: estudio de caso de la Universidade de Vigo (UVigo), en España. En AA.VV. (Eds.) *La COVID-19 llega a las aulas.* Colección Estudios Aranzadi. Thomson Reuters-Aranzadi. En prensa.

Miquel-Vergés, J. (2022b) La docencia universitaria en tiempos de pandemia. Estudio de caso. En AA.VV. (Eds.) *COVID-19: Imperativo para el cambio*. Colección Comunica. Tirant lo Blanch. En prensa.

Nelson, D. (2020, 16 de septiembre). *Game Developers Look To Voice AI For New Creative Opportunities*. Unite.AI. https://cutt.ly/iK0SRxO

Olivas, O. (2016, 3 de mayo). *¿Qué es el face swap y cómo lo usan las marcas?* Merca2.0. https://cutt.ly/NLlwPeW

Panorama Audiovisual. (2021, 9 de febrero). *Así se hizo el deepfake de Lola Flores para el spot 'Con mucho acento'.* Panorama Audiovisual. https://cutt.ly/YK687kU

Pérez Navarro et al. (2020, 27 de noviembre). El uso de vídeos en la educación puede mejorar el volumen de aprobados. *Blog del eLearning Innovation Center - UOC*. https://cutt.ly/xLrxLOF

Prajwal, K. R. et al. (2019a). *Towards Automatic Face-to-Face Translation*. *CVIT*. https://cutt.ly/jLf2AFa

Prajwal, K. R. et al. (octubre de 2019b). *Towards Automatic Face-to-Face Translation*. MM '19: Proceedings of the 27th ACM International Conference on Multimedia, octubre 2019, pp. 1428-1436. https://doi.org/10.48550/arXiv.2003.00418

Real Academia Española. (s.f.). Doblaje. En *Diccionario de la lengua española*. Recuperado el 23 de julio de 2022, de https://dle.rae.es/doblaje?m=form

Rudrabha Mukhopadhyay (2019). *Face to Face Translation* [video]. YouTube https://cutt.ly/MLk5U8L

Snapchat. (13 de abril de 2022). En *Wikipedia*. https://cutt.ly/PL2IM9x

Snow White and the Seven Dwarfs. (19 de julio de 2022). En *Wikipedia*. https://cutt.ly/bL2OrUG

Solsman, J. (2019, 12 de julio). *Deepfakes freak YouTubers out. Vidcon offers a way to prepare.* CNET. https://cutt.ly/LK0Gdic

Spotify. (Junio de 2020a). *Cómo la Inteligencia Artificial recreó la voz de Franco* (Nº EXTRA) [Episodio de Podcast]. En *XRey*. Spotify. https://cutt.ly/kK69GF1. También disponible en *XRey: el caudillo digital* [video]. YouTube https://cutt.ly/VL30MOM

Spotify. (Junio de 2020b). *El temblor* (Nº 4) [Episodio de Podcast]. En *XRey*. Spotify. https://cutt.ly/kLqsiT3

Suárez Iglesias, D. (2019). *Diccionario de anglicismos y otros extranjerismos*. Defensadelidioma. https://cutt.ly/fLrgw0e

The Dalí Museum (2019, 8 de mayo). *Behind the Scenes: Dali Lives* [vídeo]. YouTube. https://cutt.ly/MLk5SUr

Veritone. (2022) *Deepfake voice: everything you need to know about deepfake voice*. Veritone. https://cutt.ly/SK8icMm

Vicomtech (2022a). *Clonación de voz.* https://cutt.ly/yK8De99

Vicomtech (2022b, 6 de abril). *Vicomtech expone las claves de las tecnologías de IA aplicadas a la generación y detección de DeepFakes*. https://cutt.ly/lK09sbL

Vicomtech (2022c, 8 de abril). *Casos reales de Deep Fake. Generación y detección de imagen y voz sintéticas a través de IA* [vídeo]. YouTube. https://cutt.ly/dLkGE3a

Vilardi, R. (2019, 29 de agosto). *El tono de voz en branding: la siguiente frontera*. Wearemarketing. https://cutt.ly/tLlgQBO

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11): 40-53. http://doi.org/10.22215/timreview/1282

YouMoretv-Espectáculo. (2014, 16 de diciembre). *Los del Río "Macarena"* [vídeo]. YouTube. https://cutt.ly/fLwF0Z2