UNIVERSITAT POLITÈCNICA DE VALÈNCIA | IN-RED 2022 VIII Congreso de Innovación Educativa y Docencia en Red

Congreso In-Red 2022
UPV, 6 – 8 de julio de 2022
Doi: http://dx.doi.org/10.4995INRED2022.2022.15844

Doblaje automático de vídeo-charlas educativas en UPV[Media]*

Alejandro Pérez González de Martos[a], Adrià Giménez[a], Javier Jorge[a], Javier Iranzo-Sánchez[a], Joan Albert Silvestre-Cerdà[a], Gonçal V. Garcés Díaz-Munío[a], Pau Baquero-Arnal[a], Albert Sanchis[a], Jorge Civera[a], Alfons Juan[a] y Carlos Turró[b]

[a]Machine Learning and Language Processing group (MLLP), Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV)
[b]Àrea de Sistemes d'Informació i Comunicacions (ASIC), UPV

How to cite: A. Pérez-González-de-Martos, A. Giménez, J. Jorge, J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, G. V. Garcés Díaz Munío, P. Baquero-Arnal, A. Sanchis, J. Civera, A. Juan y C. Turró. 2022. Doblaje automático de vídeo-charlas educativas en UPV[Media]. En libro de actas: *VIII Congreso de Innovación Educativa y Docencia en Red*. Valencia, 6 – 8 de julio de 2022. http://dx.doi.org/10.4995/INRED2022.2022.15844

Abstract

More and more universities are banking on the production of digital contents to support online or blended learning in higher education. Over the last years, the MLLP research group has been working closely with the UPV's ASIC media services in order to enrich educational multimedia resources through the application of natural language processing technologies including automatic speech recognition, machine translation and text-to-speech. In this work we present the steps that are being followed for the comprehensive translation of these materials, specifically through (semi-)automatic dubbing by making use of state-of-the-art speaker-adaptive text-to-speech technologies.

Keywords: *automatic speech recognition, machine translation, text-to-speech, automatic dubbing, OER*

Resumen

Cada vez son más las universidades que apuestan por la producción de contenidos digitales como apoyo al aprendizaje en línea o combinado en la enseñanza superior. El grupo de investigación MLLP lleva años trabajando junto al ASIC de la UPV para enriquecer estos materiales, y particularmente su accesibilidad y oferta lingüística, haciendo uso de tecnologías del lenguaje como el reconocimiento automático del habla, la traducción automática y la síntesis de voz. En este trabajo presentamos los pasos que se

*Este trabajo ha recibido financiación del Gobierno de España a través de la subvención RTI2018-094879-B-I00 financiada por MCIN/AEI/10.13039/501100011033 (Multisub) y por "FEDER Una manera de hacer Europa"; del programa Erasmus+ Educación a través del acuerdo de subvención 20-226-093604-SCH (EXPERT); *and by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 761758* (X5gon).

están dando hacia la traducción integral de estos materiales, concretamente a través del doblaje (semi-)automático mediante sistemas de síntesis de voz adaptables al locutor.

***Keywords**: reconocimiento automático del habla, traducción automática, síntesis de voz, doblaje automático, OER*

1 Introducción

Estos últimos años hemos visto cómo la educación en línea ha jugado un papel esencial a causa de la situación pandémica global provocada por la COVID-19. A su vez, la producción digital de contenidos educativos, aunque ya en auge, se ha visto claramente potenciada por este contexto, y particularmente en el caso de la enseñanza superior. En el caso concreto de UPV[Media], el repositorio institucional de la Universitat Politècnica de València, el número de vídeos de apoyo a la docencia ha pasado de 44000 (desde su creación en 2007) a más de 88000 en el periodo comprendido entre junio de 2020 y marzo de 2022.

Los distintos catálogos audiovisuales de UPV[Media] son un ejemplo de cómo están evolucionando los repositorios educativos, no solo en cuanto a tamaño, sino también en su complejidad en términos lingüísticos. En el caso de UPV [Media], aunque la mayoría de materiales son producidos únicamente en español, la UPV apuesta por potenciar la educación multilingüe con el fin de formar estudiantes plurilingües competentes, como mínimo, en las dos lenguas oficiales de la Comunitat Valenciana (castellano y valenciano) y en la lengua considerada como internacional (inglés) (BOUPV20, 2020, pp 120–144). Para ello es importante hacer un esfuerzo en la utilización de los recursos disponibles con el objetivo de tratar los distintos idiomas de manera igualitaria. Con el fin ampliar el soporte lingüístico en UPV[Media], sus distintos catálogos han sido elegidos como casos de estudio en diversos proyectos europeos relacionados con la aplicación de tecnologías del habla en el contexto de la educación universitaria en línea, y en concreto con tecnologías de reconocimiento automático del habla (RAH) y traducción automática (TA). Entre estos proyectos se encuentran "transLectures: Transcription and Translation of Video Lectures" (Silvestre-Cerdà y col., 2012) (2012-2014), "EMMA: European Multiple MOOC Aggregator" (Brouns y col., 2015) (2014-2016) y "X5gon: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network" (Iranzo y col., 2020) (2017-2020), donde se perseguía, entre otros objetivos, la producción automática de subtítulos multilingüe de alta calidad en repositorios educativos (vídeo-charlas, MOOCs, OER, etc.) a través de la aplicación de estas tecnologías. El probado éxito de esta aproximación (Valor-Miró y col., 2015) resultó en la integración y aplicación de estas tecnologías en el entorno de producción de UPV[Media], donde desde 2014 se transcriben y se traducen automáticamente gran parte de sus contenidos, facilitando así su accesibilidad y ayudando a superar la barrera idiomática. Se ha comprobado que la producción de subtítulos multilingüe para los contenidos de ciertas plataformas educativas en línea (de tipo MOOC, o Cursos Online Masivos y Abiertos) puede ayudar a incrementar el número de alumnos matriculados en estos cursos hasta en un 70 % (Valor Miró y col., 2018).

Los distintos avances en el aprendizaje profundo basado en redes neuronales han supuesto un salto cualitativo para un gran número de tareas en el campo del aprendizaje automático, y particularmente en multitud de tareas en el ámbito del procesamiento del lenguaje natural, como son el RAH y la TA. Los sistemas de RAH y TA se han visto gradualmente impulsados por estos avances a lo largo de los últimos años, llevando las tasas de error a niveles nunca antes vistos en estas tecnologías, en algunos casos comparables a las de transcriptores o traductores humanos. El siguiente paso natural hacia la *traducción integral* (es decir, la traducción de un objeto multimedia

como si hubiese sido producido originalmente en el idioma destino) es la aplicación de sistemas estado del arte de síntesis de voz sobre los subtítulos traducidos para el doblaje automático de estos materiales (Piqueras y col., 2017). Esto supone varios retos. En primer lugar, construir sistemas capaces de producir voz sintética natural de alta calidad en cada uno de los idiomas considerados. En segundo lugar, conseguir que estos sistemas sean capaces de adaptarse a la voz del locutor en el idioma destino, de forma que la voz sintética suene lo más parecida posible al locutor del vídeo original aunque se trate de idiomas distintos. En tercer lugar, es necesario establecer algoritmos o mecanismos para la correcta sincronización entre la pista de audio sintética generada y la pista original, de forma que no se incurra en retrasos o adelantos significativos entre ambas. Por último, para lograr un resultado óptimo, es conveniente eliminar la parte de habla de la pista de audio original y recuperar el audio residual con el fin de incluir música u otros sonidos de ambiente relevantes presentes en el vídeo original también en la pista doblada.

2 Objetivos

El objetivo principal de este trabajo es dar un paso más hacia la traducción integral de los contenidos de UPV[Media] a través de la aplicación de tecnologías de síntesis de voz basadas en redes neuronales. Esto permitiría el doblaje automático (o semi-automático) de estos contenidos, facilitando por un lado la accesibilidad a estas traducciones a personas con discapacidad visual, y por otro el consumo de estos contenidos al resto de estudiantes de modo que no sea necesario dividir la atención entre los subtítulos y la presentación.

Para la consecución de este objetivo, se identifican cuatro pasos necesarios:

1. Desarrollar sistemas de síntesis de voz en castellano, valenciano e inglés capaces de adaptarse dinámicamente a la voz de cualquier locutor.
2. Diseñar un algoritmo o sistema capaz de ajustar la velocidad de habla de la voz sintética para no incurrir en retrasos significativos respecto al audio original.
3. Recuperar el audio residual de las presentaciones originales e incluirlo en la pista de audio sintetizada.
4. Hacer accesible esta tecnología a través de TLP y diseñar un flujo de trabajo para la óptima gestión del proceso de síntesis de voz desde UPV[Media].

3 Desarrollo de la innovación

Esta sección está organizada de la siguiente forma. En la Sección 3.1 se presenta UPV[Media], prestando especial atención a los aspectos lingüísticos entre sus distintos catálogos, junto con los sistemas de RAH y TA empleados para la producción automática de subtítulos multilingüe. La Sección 3.2 presenta los sistemas de síntesis de voz adaptables al locutor desarrollados para su utilización UPV[Media]. Finalmente, la Sección 3.3 describe el proceso de doblaje junto con el flujo de trabajo adoptado por UPV[Media] para la obtención de resultados de calidad óptima para su publicación.

3.1 Subtitulación multilingüe en UPV[Media]

UPV[Media] es un servicio profesional de la UPV para la creación, almacenamiento, gestión y diseminación de contenidos educativos en formato audiovisual (MediaUPV, 2020; Turró y col., 2009). Fué lanzado en 2007, e inicialmente fué concebido para facilitar a los profesores de la UPV la grabación de pequeños vídeos formativos de alta calidad en un entorno de estudio de grabación profesional. El objetivo de estos vídeos era servir de apoyo al aprendizaje combinado (en inglés, *blended learning*) a través de estas pequeñas *píldoras de conocimiento*. Estas píldoras, llamadas *poliMedias*, han servido también como base para la creación de MOOCs (*Massive Online Open Courses*) (UPVX, 2020) desde la UPV, especialmente desde que es miembro de edX (2014) (UPValenciaX, 2020). Cabe remarcar que la UPV se ha erigido como una de las instituciones con más prestigio en la creación de MOOCs en español, con más de 103 cursos, 591 ediciones, 3,4 millones de inscripciones y 8 de entre los 250 mejores cursos de todos los tiempos (ClassCentral, 2022). Además de los poliMedias, UPV[Media] ha ido incluyendo otros tipos de vídeos, como vídeos de grabación propia producidos directamente por los profesores y estudiantes de la UPV, llamados *poliTubes*, o grabaciones de clases con materiales de grabación instalados en las mismas aulas, gestionados a través de la plataforma de código abierto Opencast[1] y servidos a los estudiantes a través de Sakai LMS[2] (Opencast, 2020; Turró y col., 2014).

La Figura 1 muestra los estudios UPV[Media] dedicados a la grabación de poliMedias, compuestos fundamentalmente por un croma, una cámara de vídeo, una estación de captura, un micrófono de solapa y un sistema de iluminación básico. A través de un sistema de reservas online, el profesor acude al estudio con sus diapositivas y realiza su presentación frente a la cámara, donde se captura simultáneamente su interacción con las diapositivas, y finalmente se mezclan ambas entradas, resultando en un formato como el que se muestra en la Figura 2. La Tabla 1 muestra la cantidad de poliMedias disponibles en cada uno de los tres idiomas principales de UPV[Media]: castellano, valenciano e inglés.

Fig. 1: Estudio de grabación [UPV]Media dedicado a las grabaciones de poliMedias.

[1] https://opencast.org/
[2] https://www.sakailms.org/

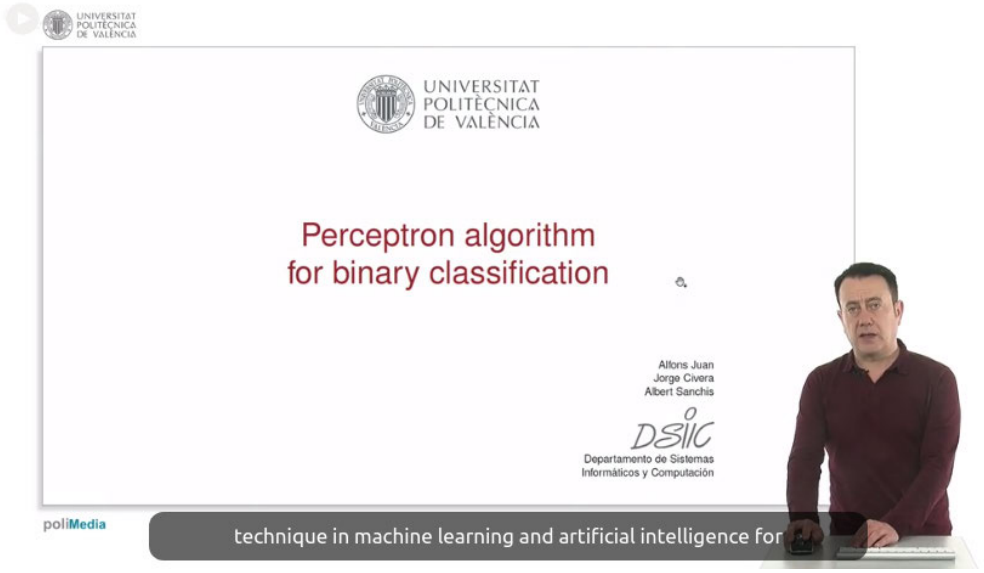

Fig. 2: Un vídeo poliMedia con subtítulos automáticos.

Tabla 1: Número de poliMedias en castellano, valenciano e inglés (Marzo 2022).

	Vídeos		Horas	
Idioma	N.º	%	N.º	%
Castellano	21259	87	3474	90
Valenciano	578	2	69	2
Inglés	2560	11	315	8
Total	24397	100	3858	100

La Figura 3 muestra un ejemplo de vídeo en formato poliTube, de producción casera y donde únicamente se muestra la pantalla correspondiente a la presentación que realiza el profesor. La Tabla 2 detalla la cantidad de vídeos y horas de vídeo correspondientes a poliTubes en castellano, valenciano e inglés.

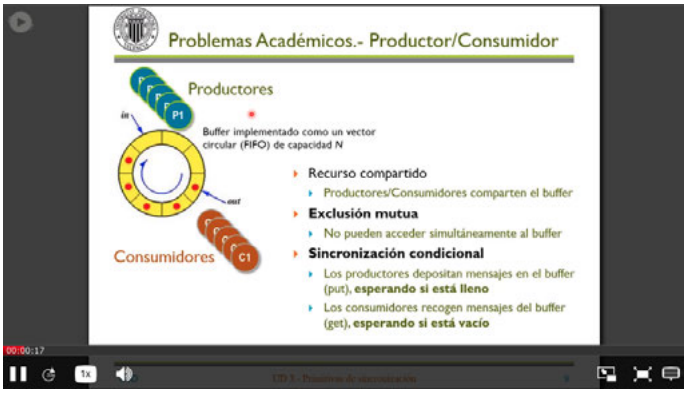

Fig. 3: Un vídeo poliTube.

Tabla 2: Número de poliTubes en castellano, valenciano e inglés (Marzo 2022).

	Vídeos		Horas	
Idioma	N.º	%	N.º	%
Castellano	38306	84	34620	76
Valenciano	1399	3	759	2
Inglés	6039	13	10314	22
Total	45744	100	45693	100

Como puede observarse, tanto los poliMedias como los poliTubes se producen principalmente en castellano, mientras que solo una pequeña parte se produce en valenciano o inglés. Cabe destacar que el número de poliMedias y poliTubes disponibles en inglés es entre cuatro y cinco veces superior al de valenciano, cuando ambos idiomas comparten una oferta académica similar (BOUPV20, 2020, pp 120–144). Esto es debido, fundamentalmente, a que los estudiantes cuya lengua materna es el valenciano son también hispanoparlantes, y por tanto los poliMedias en castellano son usados indistintamente como apoyo al aprendizaje combinado en valenciano.

Como se indicaba anteriormente, desde 2014 UPV[Media] integra la plataforma de código abierto TLP (transLectures-UPV Platform) (Pérez González de Martos y col., 2015; Silvestre-Cerdà y col., 2013) para la transcripción y traducción automática de sus contenidos a través de sistemas de RAH y TA diseñados para tal fin. Esto permite la generación automática de subtítulos en varios idiomas, entre los que se incluye siempre el idioma original (transcripción) y una serie de traducciones en base a los pares de lenguas soportados por los sistemas de TA disponibles. TLP incluye también una aplicación web para la post-edición de los subtítulos generados automáticamente (ver Figura 4). A través de este editor, el profesor puede corregir, si considera necesario, la salida de los sistemas automáticos en sus vídeos. También se posibilita la post-edición de los subtítulos a todos los alumnos de la UPV, cuyos cambios deberán ser previamente aprobados por el autor del vídeo antes de ser publicados.

Fig. 4: Aplicación web de TLP para la post-edición de subtítulos.

Los sistemas de RAH desarrollados para UPV[Media] siguen la aproximación híbrida descrita en Bourlard y Wellekens, 1990. En esta aproximación, el sistema cuenta con dos modelos estadísticos independientes: un modelo acústico y un modelo de lenguaje. Para una descripción técnica más detalla se puede consultar Jorge y col., 2021. Los datos utilizados para el entrenamiento de ambos modelos en cada uno de los tres idiomas considerados en este trabajo (castellano, valenciano e inglés) están formados por un conjunto heterogéneo proveniente de diversas fuentes. La Tabla 3 resume la cantidad de esos datos empleada para el entrenamiento de los modelos junto al tamaño del vocabulario de cada sistema.

Tabla 3: De izquierda a derecha: tamaño del vocabulario del sistema, número total de palabras utilizadas para entrenar el modelo de lenguaje y número total de horas de audio utilizadas para entrenar el modelo acústico.

	Vocabulario (K)	N.º palabras (G)	Horas
Castellano	255,5	3,4	3907,6
Catalán	323,2	2,7	2919,4
Inglés	300,0	17,9	6039,3

Para la evaluación de los sistemas de RAH se emplea la métrica conocida como ratio de error por palabra (WER, por sus siglas en inglés *Word Error Rate*), ampliamente utilizada en el área, que se puede interpretar de forma aproximada como el porcentaje de errores a nivel de palabra. Formalmente, el WER se define como:

$$\mathrm{WER} = \frac{S + D + I}{N_r} \tag{1}$$

donde N_r es el número total de palabras de la referencia, y S, D e I son respectivamente el número de sustituciones, borrados e inserciones de palabras necesarios para convertir la frase de referencia

en la salida del sistema automático. A modo orientativo, dentro del ámbito de la subtitulación asistida por ordenador, se comprueba empíricamente que una tasa de error cercana al 15 % implica un esfuerzo de subtitulación equivalente al doble de la duración del audio original, mientras que una tasa del 35 % implica un esfuerzo similar a la transcripción completa del audio desde cero (Bain y col., 2005). Por tanto, valores cercanos al 15-20 % resultan de gran utilidad para la mayoría de las tareas relacionadas con el reconocimiento del habla, mientras que valores inferiores del 10-15 % pueden ofrecer resultados realmente buenos en los que, en la mayoría de casos, pueden ser utilizados sin supervisión alguna (i.e: subtitulación de eventos en directo).

Estos mismos sistemas han sido construidos para poder ser utilizados también en reconocimiento del habla en directo con latencias de entre 0,7 y 1,0 segundos. Es decir, los sistemas son capaces de generar la transcripción de una señal de longitud ilimitada con una latencia inferior a 1s. Junto a sistemas de TA y síntesis de voz preparados también para el funcionamiento en tiempo real, se podrían construir sistemas que permitan al alumno seguir una clase que se está impartiendo en ese mismo instante en un idioma distinto al original, con tan solo unos pocos segundos de retardo. La Tabla 4 muestra los valores de WER obtenidos por cada sistema en los conjuntos de evaluación de poliMedia, extrapolables a los de otros vídeos de similares características como los poliTubes. Cabe mencionar que el MLLP ha obtenido recientemente el primer puesto en el *IberSpeech-RTVE 2020 TV Speech-to-Text Challenge*, una competición de RAH para la transcripción de programas de televisión organizada en colaboración con Radiotelevisión Española (RTVE), empleando esta misma tecnología (Baquero-Arnal y col., 2022), que la UPV también exporta a otras instituciones a través de convenios de colaboración y transferencia tecnológica, como en el caso de la televisión valenciana À Punt[3].

Tabla 4: WER % obtenidos en los conjuntos de evaluación de poliMedia (diferido y directo).

	WER % (Diferido)	WER % (Directo)
Castellano	8.3	8.7
Valenciano	11.2	11.5
Inglés	12.0	13.4

Respecto a los sistemas de TA desarrollados para UPV[Media], en Iranzo-Sánchez y col., 2021 (de corte más técnico) se describen los modelos estadísticos y procedimientos de entrenamiento empleados. Adicionalmente, estos sistemas han sido adaptados a dominios específicos (vídeos educativos en este caso), ya que la adaptación al domininio se ha demostrado efectiva en la obtención de mejoras significativas en la calidad de la traducción (Baquero-Arnal y col., 2019).

Siguiendo la aproximación descrita en Iranzo-Sánchez y col., 2021, se han desarrollado diversos sistemas para su uso en UPV[Media], salvo en el caso de la traducción entre castellano y valenciano donde se emplea el software de código abierto *Apertium*[4]. Por su relevancia en UPV[Media], a continuación se describe el sistema de TA desde castellano a inglés. Para el entrenamiento del sistema, se han utilizado 65 millones de pares de frases, obtenidas de diversos recursos de dominio público de la red OPUS-nlp (Tiedemann, 2012). Para evaluar la calidad de las traducciones se han utilizado las dos métricas comúnmente empleadas en el área para tal fin: BLEU (*BiLingual Evaluation Understudy*) (Papineni y col., 2002) y TER (*Translation Error Rate*) (Snover y col., 2006). Por regla general, valores de BLEU por encima de 35 se consideran traducciones de alta

[3] https://www.upv.es/noticias-upv/noticia-12340-accesibilidad-es.html
[4] https://www.apertium.org/

calidad (Iranzo y col., 2020; Valor Miró y col., 2018). En este caso, la calidad del sistema castellano-inglés de UPV[Media] se evalúa sobre un conjunto de *test* formado por 1139 pares de frases extraídas de poliMedia, obteniéndose valores de 35,9 BLEU y 51,1 TER.

3.2 Síntesis de voz adaptable al locutor

Durante los últimos años, la naturalidad alcanzada por los sistemas de síntesis de voz se ha visto drásticamente mejorada gracias a la adopción de tecnologías de aprendizaje profundo que ya estaban cosechando grandes éxitos en otras tareas y campos del aprendizaje automático. La naturalidad de los sistemas actuales es tal que en muchas ocasiones la voz sintética no es claramente distinguible de la voz humana (Pérez y col., 2021; Shen y col., 2018). Los sistemas de síntesis de voz basados en redes neuronales, al igual que sus homólogos de RAH y TA, se entrenan a partir de colecciones de datos etiquetados. En el caso de la síntesis de voz, estos datos se corresponden con grabaciones de audio realizadas por uno o más locutores (idealmente en un entorno controlado y en calidad de estudio) acompañadas de sus correspondientes transcripciones en cada uno de los idiomas considerados.

Con el objetivo de desarrollar sistemas de síntesis para UPV[Media], durante los cursos académicos 2016-17 y 2017-18 se llevó a cabo el registro de una base de datos de grabaciones de voz en castellano, valenciano e inglés por los profesores de la UPV a través del programa *Docència en Xarxa*. Ésta contiene un total de 36,3, 8,5 y 14,0 horas de grabaciones en castellano, valenciano e inglés, respectivamente, realizadas por un total de 98 profesores (Piqueras y col., 2017). Sin embargo, esta cantidad de datos (y de locutores) es escasa para el desarrollo de modelos capaces de adaptarse a (o imitar) la voz de cualquier locutor, incluso cuando éste no forme parte del conjunto de datos empleado para entrenar los sistemas. En este caso, para unos buenos resultados en términos de adaptación es recomendable que la colección de datos contenga el mayor número de locutores posible. Por tanto, a este conjunto de datos se le añaden otros de dominio público (Kjartansson y col., 2020; Zen y col., 2019) a la hora de entrenar los sistemas.

Los sistemas propuestos capaces de adaptarse a locutores no vistos durante el entrenamiento funcionan del siguiente modo. En primer lugar, mediante un sistema de identificación del locutor desacoplado del sistema de síntesis, se extraen las características vocales del audio de referencia en el idioma original. Este sistema está desarrollado expresamente para ser capaz de extraer las características de la voz de referencia independientemente del idioma. Posteriormente estas características vocales, codificadas como valores numéricos, se emplean para condicionar el sistema de síntesis de forma que la voz resultante se asemeje lo más posible a la del locutor original. Estos modelos, entrenados con conjuntos de datos suficientemente grandes, son capaces de imitar con cierta precisión la voz de cualquier locutor en el idioma destino.

3.3 Doblaje automático en UPV[Media]

Con el fin de permitir la generación versiones dobladas de los contenidos de UPV [Media], los sistemas de síntesis de voz descritos en el apartado anterior se integran en TLP. Una vez integrados, TLP permite la generación automática de pistas dobladas siguiendo el flujo que se muestra en la Figura 5. Como puede observarse, en primer lugar, se extrae el audio del vídeo y se envía a los sistemas de RAH y TA para generar los subtítulos multilingüe. Una vez disponibles, este mismo audio es procesado por un sistema de *speech enhancement* o realce del habla para separar la voz del resto de sonidos (música, sonido ambiente, etc.). Por un lado, la voz limpia se emplea para extraer las características vocales del locutor y condicionar la síntesis en éstas. Por otro, una vez

generada la voz sintética, el audio residual resultante se emplea para mezclar la pista de audio final traducida.

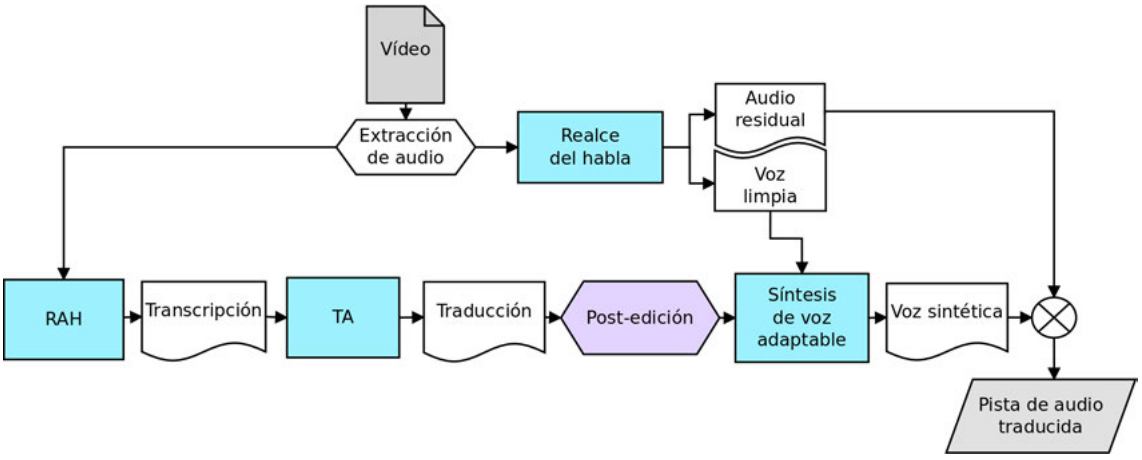

Fig. 5: Flujo completo de traducción (semi-)automática.

Con el objetivo de asegurar unos resultados de calidad publicable, UPV[Media] define un flujo de trabajo particular para el doblaje de sus contenidos. En primer lugar, la salida de los sistemas de TA se revisa manualmente para corregir posibles errores y para añadir signos de puntuación (no presentes en los sistemas actuales empleados de RAH y TA). Posteriormente, se realiza un paso de re-segmentación automática basada en esta puntuación, donde los subtítulos se dividen en frases completas en lugar del criterio estándar basado en número de caracteres por línea y silencios. Esto hace que la prosodia de la voz sintética resultante sea más natural, ya que los sistemas de síntesis se entrenan con frases completas y consecuentemente también esperan frases completas como entrada.

Por último, definimos un mecanismo de síntesis en dos pasos con el fin de adaptar la velocidad del habla sintética y evitar así incurrir en posibles retardos cuando el habla original es más rápida que ésta. En un primer paso, se obtienen las duraciones estimadas de cada segmento (frase) de la pista sintética. Seguidamente, se comparan estas duraciones con las duraciones originales de cada segmento, y se obtiene un ratio entre ambas. Finalmente, se vuelve a generar la síntesis pero indicando, para cada segmento, el ratio de velocidad de habla al que debe generarse esta síntesis, siempre dentro de un mínimo de 0,75 y un máximo de 1,35 para evitar la pérdida de naturalidad. Con este sistema se evitan, en gran medida, posibles retardos acumulados en la pista sintética respecto a la pista original.

4 Resultados

Con el fin de evaluar la naturalidad y la capacidad de adaptación de los sistemas de síntesis de voz propuestos, se lleva a cabo una evaluación subjetiva en la que 10 participantes hispanoparlantes con buen manejo del inglés evalúan ambos aspectos del sistema de síntesis en inglés. Las frases de muestra a sintetizar son extraídas aleatoriamente de la base de datos de grabaciones de *Docència en Xarxa* (no empleadas durante el entrenamiento). La evaluación consiste, por un lado, en evaluar del 1 al 5 la naturalidad de las muestras presentadas, donde unas se corresponden a muestras sintéticas y otras a muestras de control (grabaciones reales). Por otro, se evalúa también la similitud en la voz del 1 al 5 entre grabaciones reales y muestras sintéticas del mismo locutor. La Tabla 5 muestra

la puntuación de opinión media (MOS, *Mean Opinion Score*) de naturalidad con intervalos de confianza al 95 %.

Tabla 5: Naturalidad (MOS) con intervalos de confianza al 95 % (inglés).

	Naturalidad (MOS)	Muestras evaluadas
Voz sintética	$4{,}1 \pm 0{,}05$	1261
Grabaciones reales	$4{,}9 \pm 0{,}04$	387

La Tabla 6 muestra la puntuación de opinión media de similitud en la voz con intervalos de confianza al 95 %. Como puede observarse, la similitud en la voz de locutores no vistos obtiene una puntuación de 3 sobre 5, indicando que todavía hay un amplio margen de mejora sobre este aspecto en los sistemas de síntesis propuestos. Sin embargo, pensamos que la capacidad de adaptación de los sistemas es suficientemente buena como para ser empleados en este contexto.

Tabla 6: Similitud en la voz (MOS) con intervalos de confianza al 95 % (inglés).

	Similitud en la voz (MOS)	Muestras evaluadas
Voz sintética	$3{,}0 \pm 0{,}06$	1008

La integración de esta tecnología de síntesis de voz adaptable al locutor en TLP ha resultado en su implantación en UPV[Media] para el doblaje (semi-)automático de vídeos poliMedia, poliTube, MOOCs y otros tipos de vídeos. De este modo, cualquier vídeo de UPV[Media] puede ser ya doblado al castellano, valenciano o inglés con mínimo esfuerzo[5] a través de la aplicación conjunta de tecnologías de RAH, TA y síntesis de voz, facilitando así su accesibilidad y posibilitando el consumo de estos materiales en la lengua deseada, y ayudando a su vez a romper la barrera idiomática en el caso de estudiantes extranjeros. Se invita al lector a visitar el siguiente enlace[6], donde se muestra brevemente la tecnología de doblaje presentada en este trabajo.

5 Conclusiones

La traducción integral de vídeo-charlas educativas mediante tecnologías del habla innovadoras permitirá superar las barreras idiomáticas existentes en la generación y el consumo de contenidos. En este trabajo hemos presentado los pasos que el MLLP y el ASIC están llevando a cabo conjuntamente para implementar la traducción integral de los contenidos de UPV[Media]. Los sistemas presentados permiten ampliar la oferta lingüística de estos contenidos, de forma que puedan ser utilizados como apoyo a la docencia con independencia de en qué idioma han sido generados originalmente.

Los resultados resultan muy prometedores, y nos animan a seguir trabajando en esta línea, mejorando progresivamente la precisión de los sistemas de RAH y TA, y la naturalidad y capacidad de adaptación de los sistemas de síntesis de voz. También es importante seguir trabajando en la optimización de estos sistemas en términos de eficiencia computacional, reduciendo en la medida de lo posible sus requisitos hardware y el tiempo de cómputo.

[5] https://media.upv.es/#/portal/video/ca3bdb40-ac34-11ec-b8aa-4fbd1e4dbb16
[6] https://youtu.be/vd2O_n_83vI

Desde el MLLP también se está explorando la adaptación y aplicación de estas tecnologías en entornos de baja latencia, para la transcripción, traducción y doblaje (interpretación) de contenidos en directo (streaming). Ello permitiría, por ejemplo, seguir una clase en directo en un idioma distinto al que se está impartiendo, con tan solo unos pocos segundos de retardo. A su vez, esta tecnología (en concreto, la transcripción en tiempo real) también sería de utilidad para facilitar el seguimiento de una charla o clase a personas con dificultades auditivas.

Referencias bibliográficas

Bain, K., Basson, S., Faisman, A. & Kanevsky, D. (2005). Accessibility, transcription, and access everywhere. *IBM Systems Journal*, *44*, 589-604. https://doi.org/10.1147/sj.443.0589

Baquero-Arnal, P., Iranzo-Sánchez, J., Civera, J. & Juan, A. (2019). The MLLP-UPV Spanish-Portuguese and Portuguese-Spanish Machine Translation Systems for WMT19 Similar Language Translation Task (O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. L. Neves, M. Post, M. Turchi & K. Verspoor, Eds.). En O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. L. Neves, M. Post, M. Turchi & K. Verspoor (Eds.), *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, Association for Computational Linguistics. https://doi.org/10.18653/v1/w19-5423

Baquero-Arnal, P., Jorge, J., Giménez, A., Iranzo-Sánchez, J., Pérez-González-de-Martos, A., Garcés Díaz-Munío, G. V., Silvestre-Cerdà, J. A., Civera, J., Sanchis, A. & Juan, A. (2022). MLLP-VRAIN Spanish ASR Systems for the Albayzin-RTVE 2020 Speech-To-Text Challenge: Extension. *Applied Sciences*, *12*(2), 804. https://doi.org/10.3390/app12020804

BOUPV20. (2020). Official Bulletin of the UPV [Retrieved on June 2020 (in Catalan and Spanish)].

Bourlard, H. & Wellekens, C. (1990). Links between Markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(12), 1167-1178. https://doi.org/10.1109/34.62605

Brouns, F., Serrano Martínez-Santos, N., Civera, J., Kalz, M. & Juan, A. (2015, 1 de enero). Supporting language diversity of European MOOCs with the EMMA platform, En *Proc. of the European MOOC Stakeholder Summit EMOOCs 2015*, Mons (Belgium). http://www.emoocs2015.eu/node/55

ClassCentral. (2022). The Best Free Online Courses of All Time (2022) [Retrieved on March 2022].

Iranzo, J. Y col. (2020). *X5gon deliverable 3.5: Final support for Cross-lingual OER* (inf. téc.) [https://www.x5gon.org/science/deliverables]. Universitat Politècnica de València. https://www.x5gon.org/science/deliverables.

Iranzo-Sánchez, J., Jorge, J., Baquero-Arnal, P., Silvestre-Cerdà, J. A., Giménez, A., Civera, J., Sanchıés, A. & Juan, A. (2021). Streaming cascade-based speech translation leveraged by a direct segmentation model. *Neural Networks*, *142*, 303-315. https://doi.org/10.1016/j.neunet.2021.05.013

Jorge, J. Y col. (2021). Live Streaming Speech Recognition using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models [(submitted)]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Kjartansson, O., Gutkin, A., Butryna, A., Demirsahin, I. & Rivera, C. (2020). Open-Source High Quality Speech Datasets for Basque, Catalan and Galician, En *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, Marseille, France, European Language Resources association (ELRA). https://www.aclweb.org/anthology/2020.sltu-1.3

MediaUPV. (2020). The MediaUPV repository [Retrieved on June 2020].

Opencast. (2020). Opencast [Retrieved on June 2020].

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation, En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. http://www.aclweb.org/anthology/P02-1040.pdf

Pérez, A., Garcés Díaz-Munío, G., Giménez, A., Silvestre-Cerdà, J. A., Sanchis, A., Civera, J., Jiménez, M., Turró, C. & Juan, A. (2021). Towards cross-lingual voice cloning in higher education. *Engineering Applications of Artificial Intelligence*, *105*, 104413. https://doi.org/10.1016/j.engappai.2021.104413

Pérez González de Martos, A., Silvestre-Cerdà, J. A., Valor Miró, J. D., Civera, J. & Juan, A. (2015, 16 de septiembre). MLLP Transcription and Translation Platform [Short paper for demo presentation accepted at 10th European Conf. on Technology Enhanced Learning (EC-TEL 2015), Toledo (Spain), 2015.].

Piqueras, S., Pérez, A., Turró, C., Jiménez, M., Sanchis, A., Civera, J. & Juan, A. (2017, 1 de enero). Hacia la traducción integral de vídeo charlas educativas, En *Proc. of III Congreso Nacional de Innovación Educativa y Docencia en Red (IN-RED 2017)*, València (Spain). http://ocs.editorial.upv.es/index.php/INRED/INRED2017/paper/view/6812

Shen, J. Y col. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, En *Proc. of ICASSP*.

Silvestre-Cerdà, J. A. Y col. (2013). A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories, En *Proc. of 2013 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*.

Silvestre-Cerdà, J. A., Del Agua, M., Garcés, G., Gascó, G., Giménez-Pastor, A., Martínez, A., Pérez González de Martos, A., Sánchez, I., Serrano Martínez-Santos, N., Spencer, R., Valor Miró, J. D., Andrés-Ferrer, J., Civera, J., Sanchís, A. & Juan, A. (2012, 22 de noviembre). transLectures, En *Proceedings (Online) of IberSPEECH 2012*, Madrid (Spain). http://www.mllp.upv.es/wp-content/uploads/2015/04/1209IberSpeech.pdf

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation, En *Proceedings of association for machine translation in the Americas*.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS, En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Turró, C. Y col. (2009). Polimedia: a system for successful video e-learning, En *Proc. of the EUNIS Annual Congress*.

Turró, C. Y col. (2014). Deployment and Analysis of Lecture Recording in Engineering Education, En *Proc. of 2014 IEEE Frontiers in Education Conference (FIE)*.

UPValenciaX. (2020). UPValenciaX: UPV as an edX member [Retrieved on June 2020].

UPVX. (2020). UPVX: The MOOC initiative at the UPV [Retrieved on June 2020].

Valor Miró, J. D., Baquero-Arnal, P., Civera, J., Turró, C. & Juan, A. (2018). Multilingual videos for MOOCs and OER. *Journal of Educational Technology & Society*, *21*(2), 1-12. http://hdl.handle.net/10251/122577

Valor-Miró, J. D. Y col. (2015). Efficient Generation of High-Quality Multilingual Subtitles for Video Lecture Repositories, En *Proc. of the 10th European Conf. on Technology Enhanced Learning (EC-TEL)*.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z. & Wu, Y. (2019). LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech, En *Proc. Interspeech 2019*. https://doi.org/10.21437/Interspeech.2019-2441