UNIVERSITAT POLITÈCNICA DE VALÈNCIA

IN-RED 2017
Congreso Nacional de Innovación Educativa y de Docencia en Red

Congreso In-Red 2017 UPV,
13 y 14 de julio de 2017
Doi: http://dx.doi.org/10.4995/INRED2017.2017.6812

Hacia la traducción integral de vídeo charlas educativas

Santiago Piqueras[1], Alejandro Pérez[1], Carlos Turró[2], Manuel Jiménez[2], Albert Sanchis[1], Jorge Civera[1] y Alfons Juan[1]

[1]Dep. de Sistemas Informáticos y Computación, Universitat Politècnica de València
[2]Área de Sistemas de la Información y las Comunicaciones, UPV

Abstract

More and more universities and educational institutions are banking on production of technological resources for different uses in higher education. The MLLP research group has been working closely with the ASIC at UPV in order to enrich educational multimedia resources through the use of machine learning *technologies, such as automatic speech recognition, machine translation or text-to-speech synthesis. In this work, developed under the Plan de Docencia en Red 2016-17's framework, we present the application of innovative technologies in order to achive the integral translation of educational videos.*

Keywords*: multilingual, translation, MOOCs*

Resumen

Cada vez son más las universidades e instituciones educativas que apuestan por la producción de recursos tecnológicos para diversos usos en enseñanza superior. El grupo de investigación MLLP lleva años colaborando con el ASIC de la UPV con el fin de enriquecer estos materiales haciendo uso de tecnologías de machine learning, como son el reconocimiento automático del habla, la traducción automática o la síntesis de voz. En este trabajo, bajo el marco del Plan de Docencia en Red 2016-17, abordaremos la traducción integral de vídeos docentes mediante el uso de estas tecnologías.

Keywords*: multilingüe, traducción, MOOCs*

1 Introducción

En los últimos años, se ha producido un rápido crecimiento en la producción de recursos tecnológicos en la enseñanza superior y, especialmente, de material multimedia. Hoy en día son muchas las universidades e instituciones educativas que, al igual que la Universitat Politècnica de València (UPV), apuestan por la generación de vídeos docentes para su uso en MOOCs o, en general, como OER (*Open Educational Resources*) para diversos usos en enseñanza superior (p.e. en *flip teaching*).

Desde 2011, el grupo de investigación MLLP (*Machine Learning and Language Processing, www.mllp.upv.es*) y el ASIC (Área de Sistemas de la Información y las Comunicaciones) de la UPV trabajan juntos en el desarrollo de tecnologías innovadoras para la producción de vídeos docentes. Este trabajo conjunto se ha enmarcado, principalmente, de 2011 a 2014 en el proyecto europeo transLectures (Silvestre-Cerdà y col. 2012; Valor-Miró y col. 2017; transLectures 2012), de 2014 a 2016 en el proyecto europeo EMMA (Valor-Miró y col. 2017; EMMA 2017) y, desde 2016, también en el proyecto nacional MORE.

En transLectures se abordó la generación de subtítulos multilingües para repositorios de vídeos educativos mediante el uso de tecnologías de transcripción y traducción automáticas. En el marco del proyecto se generaron transcripciones y traducciones automáticas en castellano, catalán e inglés para todos los vídeos del repositorio poliMedia. Además, a través de la convocatoria 2013-14 del Plan Docencia en Red de la UPV (*DeR 2013-14*), se solicitó la colaboración de algunos profesores para la evaluación y refinamiento de las herramientas desarrolladas. Los resultados mostraron que el uso de estas tecnologías suponía una reducción sustancial en la dedicación temporal a la hora de generar subtítulos (Valor Miró y col. 2015).

Los desarrollos tecnológicos de transLectures tuvieron su continuidad en EMMA, donde se realizaron importantes avances en los sistemas de *Automatic Speech Recognition (ASR)* y *Machine Translation (MT)*. Este proyecto daba especial importancia a la producción de MOOCs por lo que, aparte de la subtitulación multilingüe de vídeos educativos, se consideró también la traducción automática de documentos de texto. Además, en EMMA se añadieron nuevas lenguas de trabajo, concretamente: italiano, francés, portugués, holandés y estonio. De nuevo, los sistemas y herramientas fueron positivamente evaluados en colaboración con el profesorado de la UPV en el marco del programa Docencia en Red 2014-15 (*DeR 2014-15*; Valor Miró y col. 2016).

Partiendo de la experiencia adquirida en transLectures y EMMA, en MORE se persigue un objetivo algo más ambicioso: incluir la *Síntesis de Voz (TTS,* del inglés *Text-To-Speech)* para abordar la traducción integral automática de OER. Esto sería, concretamente, producir una versión alternativa de un material educativo concreto (un MOOC, una vídeo-charla) como si hubiese sido creada originariamente en el idioma destino. Con el fin de atacar la traducción integral, en Docencia en Red 2016-17 (*DeR 2016-17*) se ha incorporado el registro de voz, donde el profesorado ha sido invitado a registrar grabaciones de voz en castellano, catalán e inglés.

En este documento se describe, de forma resumida, el desarrollo de sistemas de síntesis de voz adaptada en el marco de Docencia en Red 2016-17 y MORE. En la sección 2 se presentarán cuáles son los objetivos de este trabajo. En la sección 3 se detallan los tres pasos principales identificados: obtención del registro de voz del profesorado, desarrollo de herramientas de síntesis de voz e integración de las herramientas en el

flujo de procesamiento de vídeo charlas para su traducción integral. En la sección 4 se muestran los resultados obtenidos y, por último, en la sección 5 se dan algunas conclusiones y líneas de trabajo futuro relacionadas con la traducción integral de vídeo charlas docentes.

2 Objetivos

Nuestro objetivo primordial es el desarrollo de sistemas automáticos que permitan la traducción integral de vídeo charlas de UPV media. Definimos **traducción integral** de una vídeo charla como la traducción conjunta del audio e imagen (diapositivas) de la charla. El resultado de este proceso es un archivo multimedia que pueda ser visionado por el alumnado en su lengua de preferencia de manera íntegra. Es evidente que la traducción integral manual, es decir, la regrabación y retraducción de los vídeos de un repositorio como UPV media, ya sea a dos o más idiomas, es inviable. Es por ello que se propone el uso de sistemas automáticos que ayuden a reducir considerablemente el coste de este proceso.

Este trabajo está centrado en la generación de pistas de voz adaptadas al locutor que puedan ser integradas en la vídeo charla traducida. A este fin, hemos identificado tres pasos necesarios:

1. Obtención de un registro de voz multilingüe de los profesores de UPV en las tres lenguas mayoritarias en UPV media: catalán, español, inglés (Sección 3.1).
2. Desarrollo de sistemas tecnológicos innovadores que permitan la generación de voz adaptada al locutor, así como la generación de voz adaptada en lenguas que el locutor original no habla (*adaptación interlingüe*) (Sección 3.2).
3. Integración de las herramientas de síntesis en el procesado de los vídeos (Sección 3.3).

3 Desarrollo de la innovación

3.1 Registro de voz en Docencia en Red 2016/17

Uno de los principales objetivos del programa Docencia en Red 2016/17 es el de crear una base de datos, que denominaremos DeR-TTS, con registros de profesores que permita realizar traducción integral de los vídeos educativos de UPV media (*DeR 2016-17*). Este programa piloto contempla el registro de frases en catalán, español e inglés, las tres principales lenguas que ofrece UPV media. Como paso inicial, se desarrolló una interfaz web que permitiera la creación de esta base de datos (Figura 1).

Como podemos ver, esta interfaz nos muestra el usuario, idioma, frase actual y número de frases ya registradas. Mediante los controles, podemos registrar (leer) la frase actual, escuchar nuestra grabación y volverla a registrar si fuera necesario. Una vez estemos contentos con nuestro registro, apretamos el botón de aceptar, momento en el cual se almacena en la base de datos y se muestra la siguiente frase. Asimismo, se ofrece la opción de saltar frases si no se está seguro de la pronunciación y de volver a frases ya registradas si fueron aceptadas por error. Estos controles pueden ser activados mediante clicks del ratón o atajos de teclado.

Fig. 1: Interfaz de registro de voz para Docencia en Red

A cada profesor participante se le solicita que registre frases sólo en los idiomas que se considere capaz, con un mínimo de 300 frases en total, a fin de garantizar un cierto grado de calidad, y un máximo de 300 por idioma. Las frases fueron extraídas de forma semi-supervisada de distintas fuentes, tales como diarios, MOOCs o Wikipedia; y editadas si fuera necesario para facilitar su lectura.

3.2 Síntesis de voz adaptada al locutor

Los sistemas de síntesis de voz han mejorado de forma considerable en los años recientes gracias a la aplicación de técnicas avanzadas de *machine learning* y, en concreto, a las redes neuronales y al *deep learning*. Esta tecnología puntera, que ya ha dado frutos en otras áreas del procesamiento del lenguaje como ASR y MT, permite abordar problemas que quedaban fuera del alcance de las técnicas tradicionales.

En este sentido, la aproximación tradicional a la síntesis de voz requiere disponer de una base de datos de alta calidad (eso es, registrada en condiciones de estudio) y tamaño considerable para cada locutor y lengua de las cuales se quiera obtener voz sintética. Estas restricciones limitan la aplicabilidad de la síntesis de voz adaptada a repositorios educativos; dado su elevado coste y la imposibilidad de sintetizar voz adaptada a lenguas que el profesor no habla.

Para dar respuesta a esta problemática, en los últimos dos años se han desarrollado sistemas *multilingües y multi-locutor*, que permiten la adaptación de la voz sintética con menos datos e incluso sin disponer de registros del profesor en la lengua destino (adaptación interlíngüe). Estos sistemas, basados en la tecnología de redes neuronales, son capaces de combinar información procedente de registros de voz de distintos locutores y lenguas en un mismo modelo; resultando en una mejora de la naturalidad e inteligibilidad de la voz sintética. Los requisitos para una síntesis de calidad son menos estrictos que en la aproximación tradicional, dado que la red es capaz de utilizar todos los audios registrados de manera eficiente.

3.3 Traducción integral

En la Figura 2 se muestra un esquema del proceso de traducción integral que seguiría un vídeo docente, que está formado habitualmente por un vídeo grabado por el profesor, y las diapositivas sobre las que se desarrolla la charla (Figura 3). El objetivo es, por un lado, traducir el audio de la charla y, por otro, traducir los textos de la imagen (las diapositivas), todo ello de forma automática.

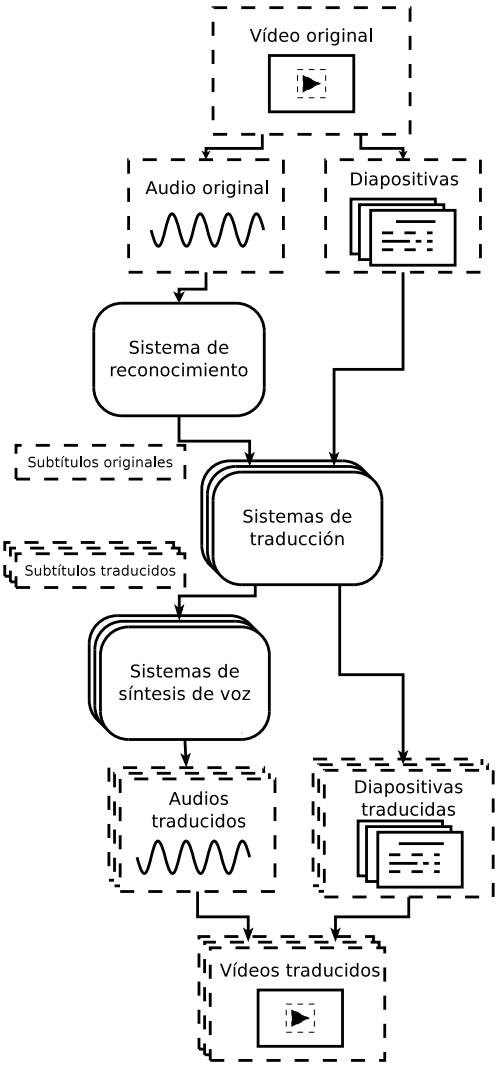

Fig. 2: Traducción integral

Para generar una pista de voz sintetizada en otro idioma, el primer paso es obtener una transcripción del audio original. Esto se consigue mediante el uso de sistemas de ASR. Una vez generada la transcripción, ésta es traducida a los idiomas destino a través de sistemas de MT. Por último, una vez tenemos subtítulos en distintas lenguas, usaremos sistemas TTS para sintetizar la voz del locutor en cada una de ellas. De forma paralela a este proceso, el texto de las diapositivas es extraído y posteriormente

traducido por los sistemas de MT. Las diapositivas traducidas, junto con las pistas de audio sintetizado, se integran para formar versiones en distintas lenguas del objeto de aprendizaje original.

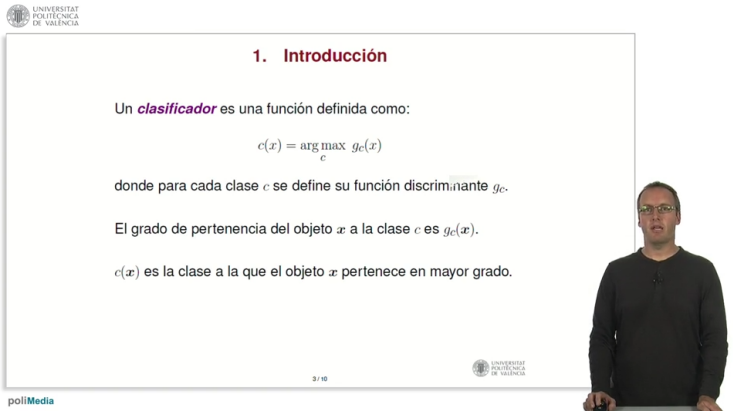

Fig. 3: Ejemplo de vídeo charla en UPV media

4 Resultados

A fecha de marzo de 2017, la base de datos DeR-TTS contiene voz de 41 locutores distintos. De estos 41, 16 han registrado su voz sólo en un idioma (14 en español), 15 han registrado su voz en dos idiomas, y 10 han registrado su voz en los tres idiomas. La Tabla 1 presenta un resumen de las estadísticas de DeR-TTS; mostrando número de locutores, número de horas y número de frases registradas por idioma.

Tabla 1: Estadísticas de la base de datos DeR-TTS. Se muestra el número de locutores número de horas y número de frases para cada uno de los tres idiomas registrados: catalán (ca), español (es) e inglés (en).

idiomas	nº loc.	duración (h)	frases (miles)
ca	19	6,3	4,5
en	19	9,7	6,5
es	38	13,9	3,8
Total	41	**29**,**9** horas	**20**,**4** frases

La acogida entre el profesorado del registro de voz como parte de Docencia en Red ha sido buena. La mayoría de participantes han registrado más de las 300 frases exigidas, y se ha dado más de un caso de participantes que han registrado todas las frases disponibles. En futuras convocatorias, queda abierta la puerta a incrementar el número de frases por idioma para dar la opción de seguir registrando voz a aquellos que así lo deseen.

Por otra parte, el desarrollo de sistemas adaptados de TTS con redes neuronales está avanzando satisfactoriamente; y ya se han obtenido primeros resultados sobre la base de datos DeR-TTS. Estos resultados confirman que los sistemas *multilingües y multi-locutor* logran un resultado superior a los sistemas tradicionales, validando así nuestra hipótesis de poder alcanzar síntesis de calidad con pocos datos por locutor.

Un análisis de corte más técnico de estos resultados se podrá encontrar en (Piqueras y col. 2017).

Por último, se han llevado a cabo tareas de integración de los sistemas de TTS dentro de la plataforma del MLLP para acercarnos más la traducción integral. A modo de demostración, en la siguiente URL se pueden encontrar algunos vídeos que han sido procesados automáticamente por los sistemas de transcripción y traducción, luego sintetizados por sistemas de voz adaptados, y finalmente alineados con las diapositivas para obtener su traducción integral:

http://mllp.upv.es/inred-17-demo/

5 Conclusiones y trabajo futuro

La traducción integral de vídeo charlas en educación superior mediante tecnologías innovadoras permitirá superar las barreras lingüísticas existentes en la generación de contenido multimedia. En este trabajo, hemos presentado los pasos que el MLLP y el ASIC están llevando a cabo para implementar la traducción integral en las vídeo charlas de UPV media. Se ha descrito el proceso de obtención de una base de datos de registros de voz, el desarrollo tecnológico de sistemas de voz adaptados al locutor y el esquema de traducción integral implementado. Los resultados resultan prometedores, y nos animan a seguir trabajando en esta línea.

El siguiente paso es estudiar la integración las herramientas de adaptación y generación de voz en la plataforma UPV media, tal y como ya se hizo con los sistemas de ASR y MT. Esta integración es esencial para medir el impacto de la traducción integral en un repositorio educativo grande como UPV media. De manera paralela, se sigue trabajando en la mejora de todos los sistemas automáticos involucrados en la traducción integral, tanto en calidad como en tiempo de procesamiento.

6 Acknowledgements

El trabajo de investigación aquí presentado ha recibido fondos del programa europeo FP7/2007-2013 en virtud del acuerdo de subvención nº 287755 (transLectures) y del ICT PSP/2007-2013 como parte del *Competitiveness and Innovation Framework Programme (CIP)* en virtud del acuerdo de subvención nº 621030 (EMMA); así como del proyecto de investigación nacional TIN2015-68326-R (MINECO/FEDER) (MORE) y de la beca VALi+d de la Generalitat Valenciana ACIF/2015/082.

Referencias bibliográficas

Docencia en Red. *DeR 2013-14*. URL: http://www.upv.es/entidades/VTIC/info/Docencia_en_Red_2013_2014.pdf.

— *DeR 2014-15*. URL: http://www.upv.es/entidades/VTIC/info/Docencia_en_Red_2013_2014.pdf.

— *DeR 2016-17*. URL: http://www.upv.es/entidades/VTIC/info/Docencia_en_Red_2016-2017.pdf.

EMMA. *Página web*. URL: https://platform.europeanmoocs.eu/.

MLLP. *www.mllp.upv.es*. URL: https://www.mllp.upv.es/.

Piqueras, Santiago y col. (2017). "Evaluation of Multi-speaker and Multilingual Long Short-Term Memory Recurrent Neural Networks for Acoustic Modeling in Speech Synthesis". En: *Proceedings of InterSpeech (sent)*.

Silvestre-Cerdà, Joan Albert y col. (2012). "transLectures". En: *Proceedings of IberSPEECH 2012*. Madrid (Spain), págs. 345-351.

transLectures. *Página web*. URL: https://www.translectures.eu.

Valor Miró, Juan Daniel y col. (2015). "Evaluación de la revisión de transcripciones y traducciones automáticas de vídeos poliMedia". En: *I Congreso Nacional de Innovación Educativa y Docencia en Red (INRED-2015)*. Universitat Politècnica de València, págs. 463-467.

— (2016). "Generación eficiente de transcripciones y traducciones automáticas en poliMedia". En: *II Congreso Nacional de Innovación Educativa y Docencia en Red (INRED-2016)*. Universitat Politècnica de València.

Valor-Miró, Juan Daniel y col. (2017). "Multilingual videos for MOOCs and OER". En: *Journal of Educational Technology & Society* (in press).