Resumen

En los últimos años, el aprendizaje profundo ha cambiado significativamente el panorama en diversas áreas del campo de la inteligencia artificial, entre las que se incluyen la visión por computador, el procesamiento del lenguaje natural, robótica o teoría de juegos. En particular, el sorprendente éxito del aprendizaje profundo en múltiples aplicaciones del campo del procesamiento del lenguaje natural tales como el reconocimiento automático del habla (ASR), la traducción automática (MT) o la síntesis de voz (TTS), ha supuesto una mejora drástica en la precisión de estos sistemas, extendiendo así su implantación a un mayor rango de aplicaciones en la vida real.

En este momento, es evidente que las tecnologías de reconocimiento automático del habla y traducción automática pueden ser empleadas para producir, de forma efectiva, subtítulos multilingües de alta calidad de contenidos audiovisuales. Esto es particularmente cierto en el contexto de los vídeos educativos, donde las condiciones acústicas son normalmente favorables para los sistemas de ASR y el discurso está gramaticalmente bien formado. Sin embargo, en el caso de TTS, aunque los sistemas basados en redes neuronales han demostrado ser capaces de sintetizar voz de un realismo y calidad sin precedentes, todavía debe comprobarse si esta tecnología está lo suficientemente madura como para mejorar la accesibilidad y la participación en el aprendizaje en línea. Además, existen diversas tareas en el campo de la síntesis de voz que todavía suponen un reto, como la clonación de voz inter-lingüe, la síntesis incremental o la adaptación *zero-shot* a nuevos locutores.

Esta tesis aborda la mejora de las prestaciones de los sistemas actuales de síntesis de voz basados en redes neuronales, así como la extensión de su aplicación en diversos escenarios, en el contexto de mejorar la accesibilidad en el aprendizaje en línea. En este sentido, este trabajo presta especial atención a la adaptación a nuevos locutores y a la clonación de voz inter-lingüe, ya que los textos a sintetizar se corresponden, en este caso, a traducciones de intervenciones originalmente en otro idioma.