Sistema de conversión de texto a habla en español con control de acento, prosodia y clonación de voz

Leonardo Pepino[1], Pablo Riera[1], Germán Barchi[1], and Joaquín Giaccio[2]

[1] Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina. {lpepino, priera, gbarchi}@dc.uba.ar
[2] Sin afiliación
joacogiaccio@gmail.com

Abstract. En esta demostración, mostraremos el sistema de conversión de texto a habla en español que hemos desarrollado, el cual permite sintetizar habla con distintos acentos latinoamericanos, controlar la prosodia y clonar y convertir voces. A su vez, el modelo es lo suficientemente liviano como para funcionar con pocos recursos de hardware, y de una calidad comparable con los principales sistemas del mercado.

Keywords: Texto a habla · Clonación de voz · Procesamiento del Habla.

Actualmente existen pocos sistemas de conversión de texto a habla (TTS) en español, y la mayoría de ellos poseen un acento neutral, de España o de México. En este proyecto enfocamos nuestros esfuerzos en desarrollar un TTS con acento argentino (rioplatense), que permita por ejemplo, una interacción más natural entre asistentes virtuales y usuarios argentinos. Además, el sistema permite elegir entre múltiples hablantes femeninos y masculinos, y múltiples nacionalidades de Latinoamérica. También es capaz de clonar voces, es decir, a partir de unos pocos minutos de habla de una persona, es capaz de sintetizar habla con la voz de esa persona. Esta característica de nuestro modelo tiene muchas aplicaciones, entre ellas, proveer a personas que perdieron o perderán su voz, una copia digital de su voz original con la que podrán comunicarse. Algunos ejemplos son pacientes con ELA o que fueron sometidos a una laringectomía, y que generalmente saben con anticipación que perderán su voz, por lo que pueden realizar grabaciones de la misma antes de perderla. Además, el sistema permite un control fino de la prosodia, pudiendo incluso ser utilizado para aplicaciones de conversión de voz. Es posible controlar la duración de los silencios, la velocidad del habla, el pitch y la energía. Por último, el modelo es lo suficientemente liviano como para ser utilizado en sistemas embebidos o con recursos de hardware limitados. A modo de ejemplo, el sistema es capaz de funcionar en una Raspberry Pi 4 con una latencia razonable. En la Figura 1 se puede observar el tiempo que tarda el sistema en sintetizar el habla respecto a la duración del habla que es sintetizada. Se puede ver que incluso sin aceleración mediante GPU, el sistema puede sintetizar habla en menos tiempo que la duración de la misma.

El modelo utilizado está basado en FastSpeech 2 [1], el cual es un TTS estado del arte que permite generar habla de forma rápida y eficiente utilizando transformers. Para evaluar la calidad de nuestro sistema, siguiendo la metodología

de evaluación de la mayoría de los trabajos de síntesis de habla [2], realizamos un test subjetivo de escucha del cual participaron 68 personas de nacionalidad argentina. En esta evaluación subjetiva, los participantes debían puntuar en una escala de 1 a 5 la naturalidad y preferencia de los audios que escuchaban, que consistían de habla sintetizada por una voz femenina y otra masculina de distintos TTS disponibles en el mercado, entre ellos el nuestro. Los resultados, mostrados en la Figura 2 indican que nuestro sistema de texto a habla (Neurasound) es competitivo contra los principales sistemas existentes en el mercado.

Por último, se pueden escuchar ejemplos en nuestro sitio web [3], e incluso utilizar el sistema en tiempo real.

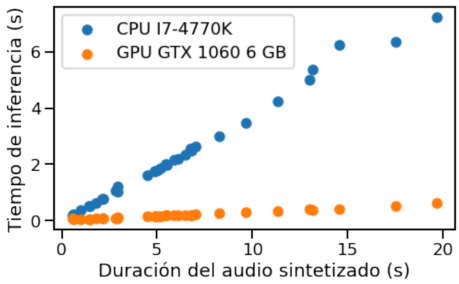

Fig. 1. Tiempo de síntesis del sistema respecto a la duración del audio sintetizado. Se muestran los resultados obtenidos en un mismo equipo, cambiando solamente el hardware en el que se realiza la inferencia (CPU vs GPU).

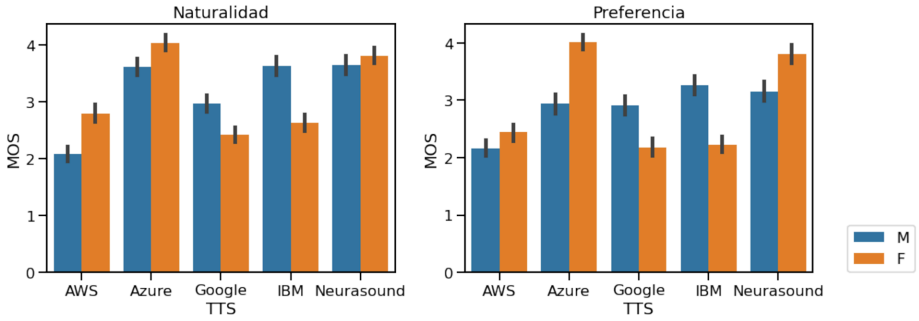

Fig. 2. Puntajes asignados a cada uno de los sistemas evaluados en las dimensiones de naturalidad y preferencia, tanto para TTS masculinos como femeninos de los principales servicios existentes en el mercado y el nuestro (Neurasound).

References

1. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T. Y.: Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv:2006.04558. (2020)
2. Cambre, J., Colnago, J., Maddock, J., Tsai, J., Kaye, J.: Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (2020)

[3] https://neurasound.com.ar/es