Conversor texto a voz multilingüe de Telefónica I+D

Ana Armenta, José Gregorio Escalada, **Juan María Garrido, Miguel Ángel Rodríguez**
Telefónica I+D
Emilio Vargas 6, MADRID 28043
aalv@tid.es, jges@tid.es, jmgarri@tid.es, miguel@tid.es

Resumen: Telefónica I+D presenta la última versión de su conversor texto a voz multiligüe, multilocutor y basado en selección de unidades
Palabras clave: CTV, multilocutor, multilingüe, síntesis por corpus

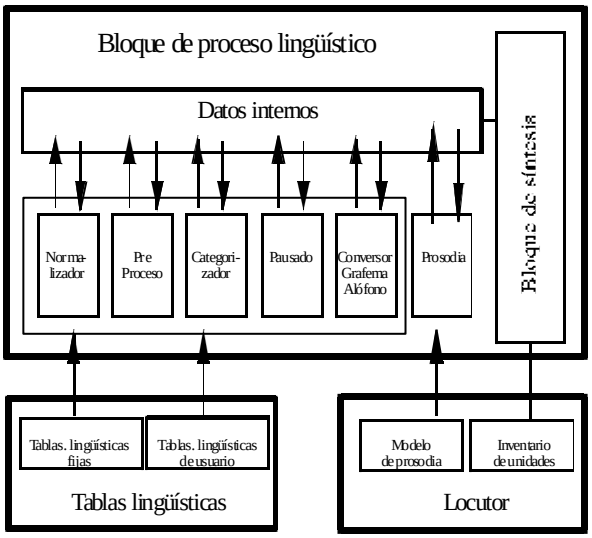
Abstract: Telefónica I+D presents the last versión its corpus based, multispeaker and multilingual text to speech system
Keywords: TTS, multiple voices, multiple languages, unit based speech synthesis

El conversor Texto-Voz (CTV) de Telefónica I+D es un sistema capaz de generar de forma automática la secuencia de sonidos que produciría una persona al leer un texto cualquiera en voz alta.

Este sistema permite convertir en voz cualquier texto escrito. La voz resultante es perfectamente inteligible y muy natural. Con el CTV, cualquier información escrita a la que pueda acceder un ordenador es leída en voz alta, y puede ser enviada a la línea telefónica o a cualquier dispositivo de salida de audio (por ejemplo, una tarjeta de audio estándar a la que se conecten unos altavoces). El CTV es un sistema multilingüe y multilocutor. Puede leer adecuadamente textos escritos en español castellano, catalán, gallego, euskera, español iberoamericano (variedad de Perú y variedad neutra, como la empleada normalmente en los doblajes de las películas), portugués de Portugal y portugués de Brasil. Para cada idioma puede generar voz sintética con locutores diferentes (por ejemplo, voz masculina y voz femenina), y es posible generar nuevos locutores basados en la voz de un locutor humano de referencia.

Los bloques principales que componen el CTV de Telefónica I+D son:

- **Proceso lingüístico**. Incorpora una serie de módulos previos (normalización, preproceso, categorización, pausado, conversión grafema-alófono) encargados de extraer información lingüística que, evidentemente, es muy dependiente del idioma de funcionamiento. También incorpora un módulo de generación de prosodia, que utiliza la información lingüística extraída por los módulos anteriores para generar los valores de los parámetros prosódicos a partir de los modelos de prosodia (específicos de cada una de las voces sintéticas del CTV).



- **Síntesis de voz**. Este bloque recibe como entrada la información lingüística procedente del bloque anterior (con los valores concretos de los sonidos que hay que generar y de los parámetros prosódicos asociados a los mismos), selecciona del

inventario los segmentos que mejor se ajustan a estos valores, y los concatena de acuerdo al modelo de síntesis empleado.

Para que el CTV esté en condiciones de generar voz sintética es preciso:

- Seleccionar las **tablas lingüísticas** correspondientes al idioma (tablas lingüísticas fijas y tablas lingüísticas de usuario), que son las que controlarán el funcionamiento del bloque de proceso lingüístico (en concreto, de los módulos previos al de generación de prosodia).
- Seleccionar un "**locuto**r" del idioma. Lo que denominamos "locutor" incluye unos modelos de prosodia y un inventario de unidades. Tanto los modelos de prosodia como el inventario de unidades se construyen a partir de unas grabaciones de la voz de un locutor humano que se toma como referencia. Los modelos de prosodia se emplean en el módulo de generación de prosodia, dentro del bloque de proceso lingüístico. El inventario de unidades se utiliza en el bloque de síntesis de voz.

Las características funcionales del CTV y sus altas prestaciones lo hacen especialmente apropiado para aplicaciones en las que no resulta práctico, o no es posible, grabar toda la información que hay que proporcionar de manera hablada. Esta situación se da cuando la cantidad de mensajes que habría que grabar es muy elevada, o cuando estos mensajes no son fijos, cambian frecuentemente o precisan una actualización constante. El CTV también resulta muy útil en aplicaciones de ayuda a discapacitados (por ejemplo, lectura de información textual para usuarios invidentes).

Algunos de los aspectos técnicos más destacados del CTV son:

- La voz sintética es generada empleando un modelo de síntesis de alta calidad, basado en la concatenación controlada de unidades.
- La selección de los segmentos a concatenar se realiza en tiempo de síntesis, de entre toda la voz almacenada en el inventario acústico.
- Se puede interrumpir la voz de salida en cualquier momento, sin necesidad de que se acabe de pronunciar el texto que esté siendo leído.
- Se pueden modificar diferentes características de la voz, como la velocidad de pronunciación, el tono y el volumen.
- Ritmo y entonación naturales, que producen un habla fluida y continua. El ritmo y la duración de los sonidos son generados de forma automática únicamente a partir del texto de entrada (sin marcas especiales).
- El CTV es capaz de procesar y expandir adecuadamente aquellos elementos del texto que no se leen directamente tal y como aparecen escritos. Por ejemplo: números cardinales, ordinales y romanos; signos especiales ("*", "#", "$"), fechas, horas, abreviaturas, acrónimos, direcciones de correo electrónico...
- Se dispone de un conjunto de tablas lingüísticas que pueden ser modificadas por los usuarios, para adaptar el funcionamiento del CTV al tipo de textos que deba leer. En estas tablas se encuentran abreviaturas, acrónimos, palabras extranjeras y signos especiales. Las tablas lingüísticas se pueden cargar y descargar dinámicamente y, para su uso, es posible seleccionar cualquiera de las cargadas.

El CTV multilingüe se utiliza con éxito en distintos servicios que ofrece el Grupo Telefónica, tales como el Portal de Voz de Telefónica Móviles, el Portal de Voz de Terra, los Servicios Administrativos de la Tarjeta Prepago de Movistar, el Servicio de Información de Consumo, aplicaciones de la Tarjeta Personal Avanzada... También lo han incorporado otras empresas a sus aplicaciones de respuesta vocal (bancos, administración pública, etc).