DESCRIPCIÓN DEL SISTEMA II DE TELEFÓNICA I+D PRESENTADO A LA EVALUACIÓN ALBAYZIN'08 PARA CTV

J. G. Escalada, A. Armenta y M. Á. Rodríguez

División de Tecnología del Habla
Telefónica Investigación y Desarrollo

RESUMEN

Se hace una descripción general del CTV Sistema II de Telefónica I+D, presentado a la evaluación de conversores texto-voz Albayzín'08. Telefónica I+D ha presentado dos sistemas a la evaluación de conversores texto-voz (denominados como Sistema I y Sistema II). Ambos sistemas comparten la mayor parte de sus componentes, y se diferencian en las técnicas de procesado de señal empleadas para la codificación, modificación y síntesis de la señal de voz. Para facilitar la visión general de cada sistema de manera independiente, la descripción de cada uno de ellos es completa, de modo que las partes comunes aparecen en ambas descripciones. Se tratan las características del sistema relacionadas con la generación de la señal de voz sintética, y el proceso de creación de la voz a partir de la base de datos con las grabaciones proporcionadas por la organización.

1. CARACTERÍSTICAS GENERALES

El Sistema II de Telefónica I+D es un CTV multilingüe y multilocutor, basado en concatenación de unidades, que emplea una técnica de selección por corpus. Este sistema emplea técnicas de programación dinámica tanto para la selección de las unidades acústicas como para la selección de las unidades entonativas

Hasta el momento, los idiomas incorporados en nuestro CTV son español castellano, catalán, gallego, euskera, portugués de Portugal, español peruano, español mexicano, español iberoamericano neutro y portugués de Brasil.

En la figura 1 se representa la estructura general de este sistema, donde se aprecian los dos bloques principales (proceso lingüístico y síntesis de voz), más una serie de tablas lingüísticas (propias del idioma de funcionamiento) y de datos acústicos y prosódicos derivados de las grabaciones de un locutor humano de referencia.

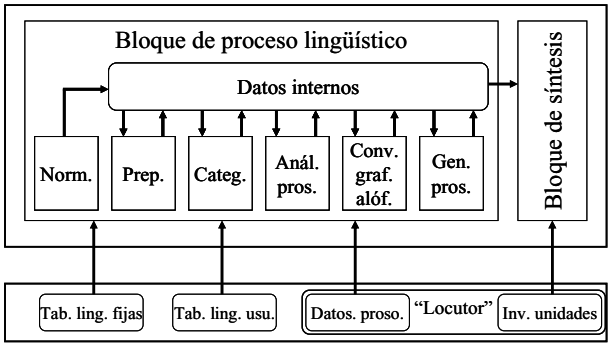

Figura 1. *Estructura del CTV Sistema II de Telefónica I+D*

2. CARACTERÍSTICAS DE LA SÍNTESIS DE VOZ

Como ya se ha dicho, el CTV Sistema II de Telefónica I+D es de los comúnmente denominados concatenativos. Genera la señal de voz sintética mediante selección y concatenación controlada de unidades acústicas.

Las unidades acústicas que maneja son difonemas que, generalmente, contienen el intervalo de señal de voz comprendido entre la parte estable de un sonido y la parte estable del siguiente sonido.

El inventario de unidades del CTV contiene multitud de opciones para cada una de las posibilidades de combinación entre dos sonidos, tantas como hayan sido incluidas en el proceso de creación de la voz. Si consideramos la combinación de sonidos a-b, el inventario contiene todos los difonemas correspondientes a las variantes de esa combinación de sonidos que aparecen en las grabaciones, y que se pueden distinguir entre sí por otras características como su contexto fonético, su F0, su duración, su localización dentro de la cadena hablada, su localización dentro de la palabra...

Para sintetizar un enunciado concreto, se hace una selección de difonemas mediante un procedimiento basado en corpus. El proceso lingüístico implementado dentro del CTV, que incorpora entre otros el módulo de análisis prosódico y el de generación de parámetros prosódicos (duraciones y contornos de F0), trata el texto para determinar cuál es la secuencia de sonidos que hay que generar, y les asigna a cada uno unos vectores de características (etiquetas) asociadas.

Con ello, se determina un "objetivo" para la síntesis: una secuencia de difonemas con características. El procedimiento de selección escoge la secuencia de difonemas recogida en el inventario de unidades que mejor se aproxima a la secuencia objetivo obtenida a partir del texto. Esta idea se concreta en un algoritmo de programación dinámica tipo Viterbi, que considera una serie de funciones de coste. La secuencia óptima es la que proporciona el coste mínimo.

Como ya se ha dicho, entre las características que se tienen en cuenta para la selección de unidades, aparte de la identidad de los sonidos implicados y su contexto, se incluyen otras de tipo prosódico. Seguidamente, describimos la forma en que se obtiene la información prosódica en nuestro sistema.

El módulo de análisis prosódico se ocupa de predecir y caracterizar los límites prosódicos en la lectura de un texto. Los límites tratados son tanto pausas (ortográficas o no) como frases entonativas, y se emplean para mejorar la generación de otros parámetros prosódicos (duración de los sonidos y contorno de F0). El funcionamiento del módulo de análisis prosódico no sólo tiene en cuenta características lingüísticas generales propias de un idioma determinado, sino que también se adapta al modo particular de hablar de un locutor humano de referencia. Este módulo ha sido personalizado usando las grabaciones suministradas para la construcción de la voz. Dentro del programa de las V Jornadas de Tecnología del Habla se presenta un artículo que describe este módulo ("Nuevo módulo de análisis prosódico del conversor texto-voz multilingüe de Telefónica I+D").

El modelo de duraciones de los sonidos es un modelo estadístico multiplicativo, cuyos parámetros se calculan para ajustarse a las duraciones recogidas en una base de datos de sonidos segmentados. Este modelo ha sido construido para la voz suministrada.

La generación del contorno de F0 se hace también mediante un procedimiento de selección por corpus de unidades entonativas elementales (patrones de F0). A partir del conjunto de patrones de F0 del corpus grabado, se compone la cadena de patrones más adecuada para construir el contorno de F0 correspondiente a las características obtenidas a partir del texto de entrada. Las unidades consideradas para la construcción de los contornos de F0 son los grupos acentuales (conjunto de sílabas comprendido entre el inicio de una sílaba tónica y el inicio de la siguiente tónica). El inventario de grupos acentuales también se obtuvo y etiquetó sobre los datos de la voz suministrada.

Las unidades acústicas almacenadas en el inventario se encuentran codificadas de acuerdo a un modelo de solapamiento y suma de ventanas (tramas) de la señal, basado en el dominio del tiempo. Es un modelo de los denominados OLA ("overlap and add") [1] que precisa conocer los instantes de tiempo de cada periodo en las zonas sonoras de la señal, en los que se centran las ventanas de análisis ("onsets" o "epochs"). En las zonas sordas, se toman ventanas a un intervalo fijo de 5 mseg. Para la determinación de estos instantes, se parte de la información obtenida por una herramienta de análisis sinusoidal de la voz (semejante a la descrita para el análisis de la voz en el Sistema I). Esta información es posteriormente filtrada y ajustada por otra herramienta que proporciona los "epochs" adecuados para el análisis OLA (buscando una adecuada localización de ventanas en las transiciones sonoro-sordo y sordo-sonoro).

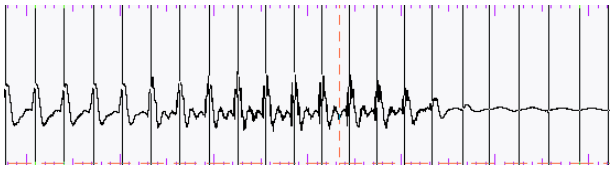

Figura 2. *Instantes de localización de las ventanas de análisis*

Una vez seleccionadas las unidades acústicas, se manipulan las tramas de voz codificada correspondientes con los propósitos siguientes:

- Modificar la duración y la entonación de los sonidos, en caso de que la diferencia entre los valores objetivo y los seleccionados supere determinados umbrales.
- Interpolar las tramas en los puntos de pegado (puntos en los que los difonemas en cuestión no se encontraban adyacentes en las grabaciones originales). La interpolación se limita a los valores de amplitud de las tramas.

El modelo de solapamiento y suma es relativamente simple y exige poca carga de cálculo, si bien no es tan flexible y robusto como el modelo sinusoidal de nuestro Sistema I en cuanto a hacer interpolaciones y modificaciones prosódicas. Cuando hay pocos pegados y el contorno de F0 se ajusta bien a las unidades seleccionadas, la calidad acústica es muy destacable.

Cuando ya se han realizado las modificaciones necesarias en la secuencia de tramas de voz codificada, se efectúa la decodificación combinando las muestras de ventanas consecutivas, para obtener las muestras de voz sintética.

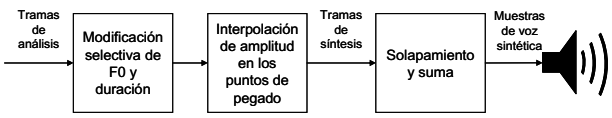

Figura 3. *Tratamiento de las tramas de las unidades acústicas seleccionadas*

3. PROCESO DE CREACIÓN DE LA VOZ

El primer paso para la creación de nuestro locutor sintético fue el tratamiento de los ficheros de texto por parte del proceso lingüístico de nuestro CTV. Ello nos

permite obtener los datos necesarios para muchas otras de las tareas implicadas en la construcción, como la transcripción fonética, el etiquetado de características lingüísticas asociadas a los sonidos...

Se obtuvieron los contornos de F0 correspondientes a los ficheros de voz, que también son necesarios como información de entrada para la codificación de la voz, el etiquetado de características asociadas a los sonidos, la construcción del inventario de grupos acentuales...

Con los contornos de F0, se hizo el análisis de tipo sinusoidal al que nos hemos referido anteriormente, para obtener la información de instantes de tiempo en los que se centran las ventanas de análisis OLA, seguido del adecuado filtrado y ajuste. No hemos empleado los instantes de tiempo proporcionados en la base de datos de voz suministrada, para mantener la compatibilidad de tipo de información y formatos con el resto de nuestras herramientas de construcción de locutores.

Acto seguido, se realizó el análisis de solapamiento y suma, determinando las ventanas de análisis sobre la señal de voz a partir de la información de los "onsets". Se emplean ventanas tipo Hanning.

Después se realizó la segmentación en alófonos de los ficheros de voz, usando nuestras propias herramientas. Aunque podríamos haber adaptado el formato de la segmentación proporcionada por la organización (para que luego fuera válido como entrada al resto de herramientas implicadas en el proceso de construcción del locutor) resultaba para nosotros más directo emplear nuestras propias herramientas de segmentación. Nuestro segmentador (descrito en [2]) se basa en hacer reconocimiento forzado mediante HMM's, y en aplicar un conjunto de reglas de lógica difusa para el ajuste posterior de la segmentación proporcionada por el reconocedor.

Una vez segmentada y etiquetada la voz por procedimientos completamente automáticos, se procedió a la construcción de los datos del módulo de análisis prosódico que son dependientes del locutor, al cálculo de los parámetros del modelo de duraciones, y a la construcción del inventario de grupos acentuales manejado por el generador de contornos de F0.

A continuación, se hizo una primera construcción del inventario de unidades acústicas, y se obtuvo un locutor de partida.

Para la construcción de este primer locutor se emplearon todas las grabaciones disponibles.

Durante el proceso de construcción, nuestras herramientas nos permiten detectar puntos en los que puede haber algún problema o desajuste. Son puntos sospechosos, en los que puede haber algún tipo de error o no: sonidos de duración excesivamente corta o excesivamente larga, desajustes en la transcripción, valores de F0 llamativos...

La localización de esos puntos sospechosos nos permite hacer un repaso selectivo de porciones de los ficheros de voz y, en caso necesario, realizar las correcciones oportunas mediante herramientas semiautomáticas. Las correcciones pueden afectar a la segmentación, a los contornos de F0 o a cualquier otro aspecto del etiquetado. Evidentemente, es mucho mejor realizar un repaso exhaustivo de todos los ficheros de voz y su etiquetado, en toda su extensión. Pero cuando esto no es posible, el repaso selectivo de porciones de los ficheros ayuda a mejorar los resultados en un plazo más corto.

Dado el tiempo limitado del que se dispuso para la construcción de la voz, pudimos hacer este repaso selectivo a una parte de los ficheros: todos los ficheros de la parte "phonetically balanced sentences" (506 ficheros), más los 172 primeros ficheros de la parte "phonetically balanced paragraphs". Con este conjunto de 678 ficheros se hizo una nueva iteración de construcción de los datos relacionados con la prosodia y del inventario de unidades, y se obtuvo el locutor con el que se hizo la generación de los estímulos de voz sintética con los textos de prueba enviados por la organización.

Los elementos componentes del locutor sintético resultante fueron los siguientes:

- Datos necesarios para el procedimiento de determinación de límites prosódicos, empleados por el módulo de análisis prosódico.
- Parámetros del modelo de duraciones de los sonidos.
- Inventario de grupos acentuales para la construcción de los contornos de F0.
- Inventario de difonemas.

El inventario de grupos acentuales contiene 4.940 elementos. De ellos, la gran mayoría corresponden a grupos extraídos de frases de modalidad enunciativa (4.707). Del resto, 212 pertenecen a frases de modalidad interrogativa, y únicamente 21 a frases de modalidad exclamativa.

En cuanto al inventario de difonemas, contiene un total de 38.004 unidades, que contienen 420 identidades de difonemas distintas (considerando que la identidad viene dada por la etiqueta de los sonidos inicial y final del difonema). Las variantes de cada identidad varían en número, desde las 780 de la unidad más frecuente [D-e] hasta el caso de identidades con una sola variante (hay 22 identidades con una sola variante).

4. BIBLIOGRAFÍA

[1] E. Moulines, F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", Speech Communication 9, pp 453-467, 1990.

[2] D. Torre, M. Á. Rodríguez, J. G. Escalada, "Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-correction Rules", Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis, pp. 207-212, noviembre 1998.