CONVERSIÓN DE VOZ CON MUY POCOS DATOS EN EVALUACIÓN AUTOMÁTICA

Daniel Erro y Asunción Moreno
Universidad Politécnica de Cataluña, Barcelona

RESUMEN

En el presente informe se describe el sistema OGMIOS Jr. sometido a la evaluación de síntesis Albayzin'06. Para mimetizar la voz de diferentes locutores se ha recurrido a la colocación de un sistema de conversión de voz a la salida de un sintetizador de voz a partir de texto. La conversión de voz se realiza por medio de una función de transformación espectral basada en información acústica, así como un escalado de la frecuencia fundamental. Se emplea el modelo de señal armónico/estocástico para la generación de la forma de onda. El sistema se testea bajo condiciones de muy poco entrenamiento en una aplicación de reconocimiento automático de locutor.

1. INTRODUCCIÓN

El propósito de los sistemas de conversión de voz es modificar la voz de un hablante *fuente* de manera que sea percibida como si se tratara de otro hablante *objetivo*, reemplazando las características físicas de la voz sin alterar el mensaje. Las técnicas de conversión de voz encuentran importantes aplicaciones en el campo de la síntesis de habla. En un sintetizador, el habla se genera habitualmente mediante la concatenación de unidades seleccionadas dentro de una base de datos, la cual ha sido construida con anterioridad grabando fragmentos de voz de un locutor cualificado. La inclusión de un sistema de conversión de voz a la salida del sintetizador permite personalizar el sistema sin necesidad de grabar bases de datos de voz de cada uno de los potenciales usuarios. Para ello simplemente se ha de entrenar una función de conversión entre la voz estándar del sintetizador, que será la *fuente*, de la que se dispone de una base de datos completa, y la voz del hablante *objetivo*, del cual se poseen en general pocos datos.

En este trabajo [1] se ha utilizado un sistema de síntesis por concatenación de unidades basado en el modelo armónico/estocástico. A su salida se ha colocado un módulo de conversión de voz. Se ha comprobado que el rendimiento global del sistema es satisfactorio aun cuando no se dispone de corpus paralelo de entrenamiento. En la presente evaluación se pretende comprobar el rendimiento del sistema en unas condiciones en las que se dispone de muy pocos segundos de entrenamiento del locutor objetivo, en torno a 10seg. En la sección 2 se describe el sistema empleado. En la sección 3 se explica el modo en que se ha trabajado con los datos para la evaluación.

2. DESCRIPCIÓN DEL SISTEMA

Se ha empleado el sintetizador de voz a partir de texto de la UPC, OGMIOS [2], que se basa en la concatenación de unidades procedentes de una base de datos. Aunque OGMIOS reconstruye la señal mediante la técnica TD-PSOLA, el bloque de generación de forma de onda ha sido reemplazado por otro basado en el modelo armónico/estocástico, que proporciona una gran flexibilidad y capacidad de modificación de la señal. El bloque de conversión de voz asociado al nuevo generador de onda se basa exclusivamente en información acústica. La figura 1 muestra el esquema del sistema global empleado. A continuación se describen brevemente los dos bloques añadidos. Para una información más detallada, consúltese [1].

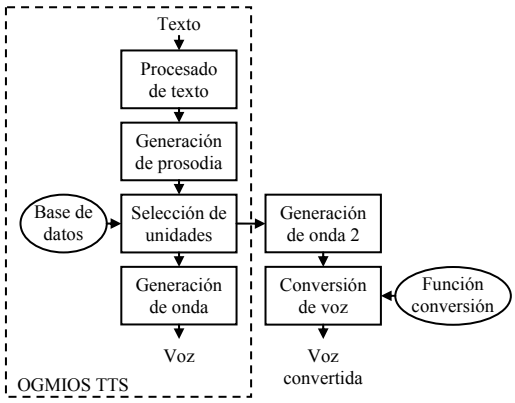

Figura 1 – Esquema del sistema

2.1. Generación de onda

El modelo armónico/estocástico representa la señal como la suma de dos componentes:

- **Armónico**: se encuentra solamente en los fragmentos sonoros, y se trata de un sumatorio de sinusoides de frecuencia múltiplo entero del pitch. Se parametriza en cada trama por medio de las amplitudes, frecuencias y fases de los armónicos situados entre 0 y 5KHz.
- **Estocástico**: se trata del residuo resultante tras eliminar de la señal todos aquellos componentes de naturaleza sinusoidal. Se modela por medio de un filtro todo-polos LPC. Se encuentra tanto en los sonidos sordos como sonoros.

Las unidades que componen la base de datos de síntesis se analizan por tramas de longitud fija N=128 muestras, correspondientes a 8ms a una frecuencia de muestreo de 16KHz. La reconstrucción de la señal se realiza

solapando tramas con una ventana triangular, de la forma siguiente:

$$s^{(k)}[n] = \sum_{j=1}^{J^{(k)}} A_j^{(k)} \cos\left(2\pi j \frac{f_0^{(k)}}{f_s} n + \varphi_j^{(k)}\right) + \sigma[n] * h_{LPC}^{(k)}[n] \quad (1)$$

$$s[kN+m] = \left(\frac{N-m}{N}\right) \cdot s^{(k)}[m] + \left(\frac{m}{N}\right) \cdot s^{(k+1)}[m-N] \quad (2)$$

para m entre 0 y N–1, siendo k el número de trama entre 1 y el número total de tramas. $A_j^{(k)}$ y $\varphi_j^{(k)}$ son las amplitudes y fases del armónico j-ésimo en la trama k. σ es un ruido blanco gaussiano y $h_{LPC}^{(k)}$ es la respuesta del filtro LPC en k. La señal reconstruida es prácticamente indistinguible de la original.

El sistema de síntesis impone unas especificaciones prosódicas concretas a cada unidad y las concatena para dar lugar a la voz de salida. Se han implementado métodos para modificar prosódicamente las señales parametrizadas de la manera descrita:

- Modificación en duración: para realizarla se modifica la separación N entre los centros de cada dos tramas. Para no romper la coherencia de la señal, se corrigen las fases de los armónicos con un término lineal.
- Modificación en pitch: las amplitudes de los nuevos armónicos se obtienen interpolando linealmente entre las anteriores, medidas en dB. Para las fases, se elimina el término lineal de la fase inicial, se interpolan las amplitudes complejas en los nuevos armónicos, y se repone el término lineal adecuado.

Para concatenar las diferentes unidades, se modifican los términos lineales de fase en las tramas de modo que haya continuidad en la forma de onda de una unidad a otra. También se suavizan las envolventes de amplitud en las fronteras de modo que no haya discontinuidades espectrales importantes. Estos aspectos, así como los relacionados con la modificación prosódica, no son evidentes, y se recomienda acudir a [1] para una explicación más detallada.

2.2. Conversión de voz

La conversión se realiza a través de las siguientes etapas [1]:

- **Normalización de pitch:** se reemplaza la media y varianza de la logf_0 del hablante origen por las del hablante objetivo, estimadas a partir de los datos de entrenamiento.
- **Conversión de tracto vocal:** a partir de las amplitudes de los armónicos se obtiene un filtro todo-polos de orden 14 que representa el tracto vocal, y se traduce a parámetros LSF en cada trama. Los LSF del hablante objetivo se obtienen aplicando una función de transformación basada en un modelo de mezclas gaussianas (GMM) de orden 8, y una vez calculados se muestrea la nueva envolvente espectral en busca de las nuevas amplitudes.
- **Estimación de envolvente de fase:** la envolvente de fase se estima a partir de la fase mínima del filtro correspondiente al tracto vocal ya transformado.
- **Predicción de componente estocástica:** se usa una función de transformación similar a la del tracto vocal, previamente entrenada, para predecir los parámetros LSF de la parte estocástica a partir de los de la parte armónica, todo ello para el hablante objetivo. En los sonidos sordos, la parte estocástica del hablante origen se mantiene sin modificación alguna.

3. EVALUACIÓN

Se pretende evaluar el sistema global con ayuda de un reconocedor de locutor. La voz base del sintetizador se ha obtenido grabando unas 10 horas de audio de un locutor profesional. Se ha preferido una voz base de hombre por proporcionar más información armónica al sistema, y por tanto mayor resolución espectral, debido a que se caracteriza por un pitch más bajo.

Sobre los hablantes-objetivo, 10 hombres y 10 mujeres, se dispone de entre 10 y 15 segundos de audio para el entrenamiento. Dado que se necesita un corpus de entrenamiento paralelo para aprender la función óptima de transformación, se han generado con ayuda del sintetizador las mismas frases pronunciadas por los hablantes-objetivo. Para alinear los vectores acústicos de manera precisa, se ha hecho uso de herramientas de segmentación basadas en modelos ocultos de Harkov (HMM) [3]. Esto da lugar a un problema: no se dispone de suficientes datos para entrenar dichos modelos con los datos de cada hablante objetivo. Es por eso que se ha recurrido a utilizar modelos ya entrenados sobre grandes bases de datos de otros hablantes, y las segmentaciones resultantes han sido corregidas manualmente.

Se ha entrenado una función de conversión para cada hablante objetivo. Por último, se han generado con el sintetizador las frases indicadas para evaluación, siéndoles aplicadas las diferentes funciones de conversión. Las señales sintéticas han sido obtenidas con muy pocos datos de entrenamiento y por tanto la calidad acústica de las mismas muestra serias degradaciones.

4. AGRADECIMIENTOS

Este trabajo se ha llevado parcialmente a cabo con fondos procedentes del proyecto TC-STAR, (FP6-506738). Se desea agradecer a Jordi Adell su gran ayuda en tareas de segmentación y síntesis.

5. REFERENCIAS

[1] D. Erro, A. Moreno, “Sistema de Síntesis Armónico/Estocástico en Modo Pitch-Asíncrono aplicado a Conversión de Voz”, IV Jornadas en Tecnología del Habla, Zaragoza, noviembre 2006.

[2] A. Bonafonte, P. D. Agüero, J. Adell, J. Pérez, A. Moreno, “Ogmios: The UPC Text-to-Speech Synthesis System for Spoken Translation”, TC-Star Workshop on Speech-to-Speech Translation, Barcelona, junio 2006.

[3] J. Adell, A. Bonafonte, J. A. Gómez, M. J. Castro, “Comparative study of Automatic Phone Segmentation methods for TTS”, ICASSP, 2005.