Práctica 5

Reconocimiento y síntesis de voz

5.1. Introducción

En esta práctica nos centraremos fundamentalmente en la síntesis de voz. En primer lugar profundizaremos en el estudio de los sintetizadores de formantes, técnica habitual de síntesis hasta mediados de los noventa, y, posteriormente, estudiaremos detalles del algoritmo de modificación prosódica TD-PSOLA, algoritmo muy utilizado hoy en día en los sistemas de síntesis por concatenación de unidades. Sin embargo, la práctica comienza con una pequeña introducción a la técnica DTW (Dynamic Time Warping), mostrando su funcionamiento y proponiendo su utilización para crear un reconocedor de dígitos extremadamente sencillo (y poco fiable). La técnica DTW permite alinear temporalmente dos secuencias de vectores y tiene también aplicación en otros campos, síntesis de voz incluido.

Además de Matlab, en esta práctica utilizaremos el programa Praat.

5.2. Dynamic Time Warping (DTW)

En este apartado mostraremos el funcionamiento de la técnica DTW para alinear dos secuencias de vectores procedentes de la parametrización de señales de voz y medir el parecido entre esas dos secuencias de acuerdo con alguna medida de distancia.

Ejercicio 1 *Los programas Matlab "DTW1" y "DTW2" muestran el funcionamiento de DTW utilizando como parametrización el módulo de la FFT de cada trama. Ejecútalos y analiza los resultados.¿Qué medida de distancia se utiliza?*

Ejercicio 2 *El programa Matlab "DTW3" muestra el funcionamiento de DTW utilizando como parametrización los coeficientes mel-cepstrum de cada trama. Ejecútalo y analiza los resultados.*

Ejercicio 3 *En este ejercicio crearemos un reconocedor de dígitos aislado muy simple. Con este objetivo crea un programa en Matlab que, tomando como referencia una única realización de cada dígito del 0 al 9, permita decidir qué dígito contiene un fichero de sonido determinado.*

5.3. Estimación de formantes

Ya hemos visto que podemos considerar el tracto vocal como una serie de sistemas de segundo orden (con polos conjugados) en cascada, es decir, su función de transferencia será

$$V(z) = \frac{\prod_{i=1}^{n}(1-z_i)(1-z_i^*)}{\prod_{i=1}^{n}(z-z_i)(z-z_i^*)} \tag{5.1}$$

siendo n el número de sistemas y z_i las raíces del denominador del filtro LPC (polos), es decir, del polinomio

$$\sum_{k=1}^{p} a_k z^{-k} = 1 \tag{5.2}$$

- Considerando que en un ancho de banda de 4 KHz hay, a lo sumo, 5 formantes, el orden de predicción, p, utilizado suele ser 10, de forma que cada sistema (par de polos conjugados) normalmente se corresponde con una resonancia del tracto vocal.

- Si un polo es real o produce un pico muy pequeño en la envolvente espectral, se considera ocasionado por la excitación y se descarta. El pico espectral (altura del formante) ocasionado por el k-ésimo par de polos conjugados será

$$A_k = \left| \frac{(1-z_k)(1-z_k^*)}{(z-z_k)(z-z_k^*)} \right|^2 \tag{5.3}$$

  siendo

$$z = e^{j2\pi F_k T} \tag{5.4}$$

$$z_k = |z_k|\, e^{j2\pi F_k T} \tag{5.5}$$

  y T el período de muestreo.

- De la expresión 5.5 se deduce que las frecuencias de los formantes, F_k, se obtienen directamente a partir de la fase de los polos mediante

$$F_k = \frac{1}{2\pi T} \arg[z_k] \tag{5.6}$$

- Además, se puede demostrar fácilmente que el ancho de banda de 3dB del k-ésimo formante viene dado por

$$B_k = -\frac{1}{\pi T} \ln |z_k| \tag{5.7}$$

Ejercicio 4 *En el fichero "estima_formantes.m"'se estiman y representan los formantes de las distintas tramas de un fichero de voz a 8 kHz. En dicho cálculo se emplea la función "lpc2formant" que debes completar. Una vez hayas completado la función ejecuta el fichero principal.*

5.4. Pitch Synchronous Overlap Add (PSOLA)

Las técnicas PSOLA son una familia de algoritmos de modificación prosódica (frecuencia fundamental y duración) entre los que destaca su variante en el dominio temporal Time-Domain PSOLA o TD-PSOLA por su sencillez y posible alta calidad de la voz resultante. No obstante esta calidad depende de un marcado más o menos preciso de los distintos periodos de la forma de onda durante los tramos sonoros. Estas marcas denominadas "marcas de pitch" se sitúan habitualmente en el máximo de cada periodo (o un punto próximo a él de forma consistente). Durante los sonidos sordos se puede considerar que las marcas de pitch están equiespaciadas (cada 10 ms, por ejemplo).

El funcionamiento del algoritmo TD-PSOLA es simple:

- En primer lugar se segmenta la señal original en una serie de segmentos enventanados de duración unas dos veces el periodo local y centrados en cada una de las marcas de pitch.
- La modificación de la frecuencia fundamental se realiza a partir de unas marcas de pitch de síntesis que están dispuestas de acuerdo con el periodo fundamental local deseado. En estas nuevas marcas se centrarán los segmentos enventanados antes de proceder a su suma solapada.
- Es evidente que una modificación de la frecuencia fundamental lleva implícita una modificación de la duración (por ejemplo, si la frecuencia fundamental se incrementa, el periodo se reduce y la longitud de la señal se acorta).
- La manipulación de la duración se consigue eliminando o repitiendo alguno de los segmentos enventanados de acuerdo con la modificación deseada. Es como si a cada marca de pitch de partida le pudieramos hacer corresponder una, varias o ninguna marca de síntesis Por ejemplo, si queremos mutiplicar la duración por un factor 1.5 (sin modificar la frecuencia fundamental), basta con que repitamos una de cada dos segmentos.

En las figuras 5.1, 5.2, 5.1 y 5.4, obtenidas de [Taylor, 2008], se ilustra el funcionamiento de este algoritmo.

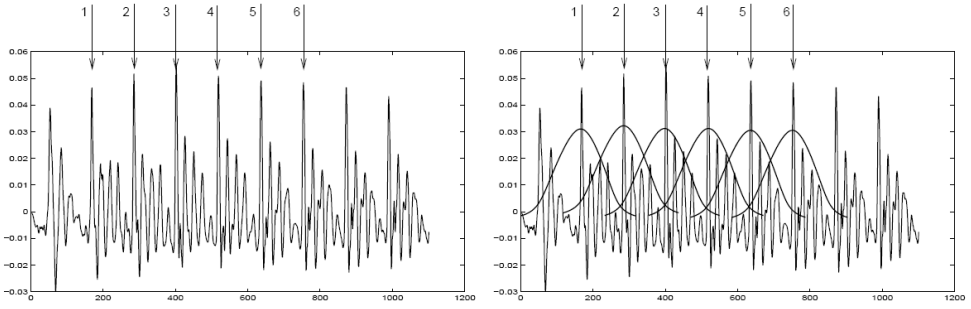

Figura 5.1: Marcas de pitch y enventanado

Ejercicio 5 *La función "pitch_marking " implementa un algoritmo para establecer las marcas de pitch. Ejecútalo con el fichero "iago800_0001.wav". Observa el resultado.*

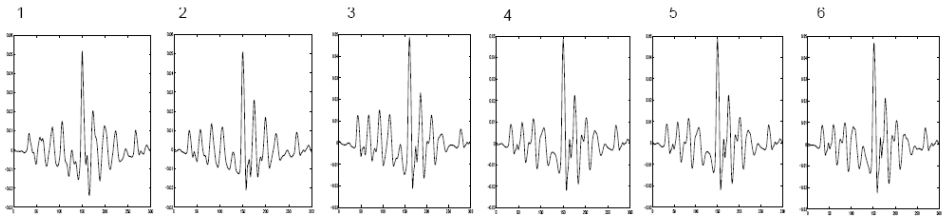

Figura 5.2: Segmentos enventanados

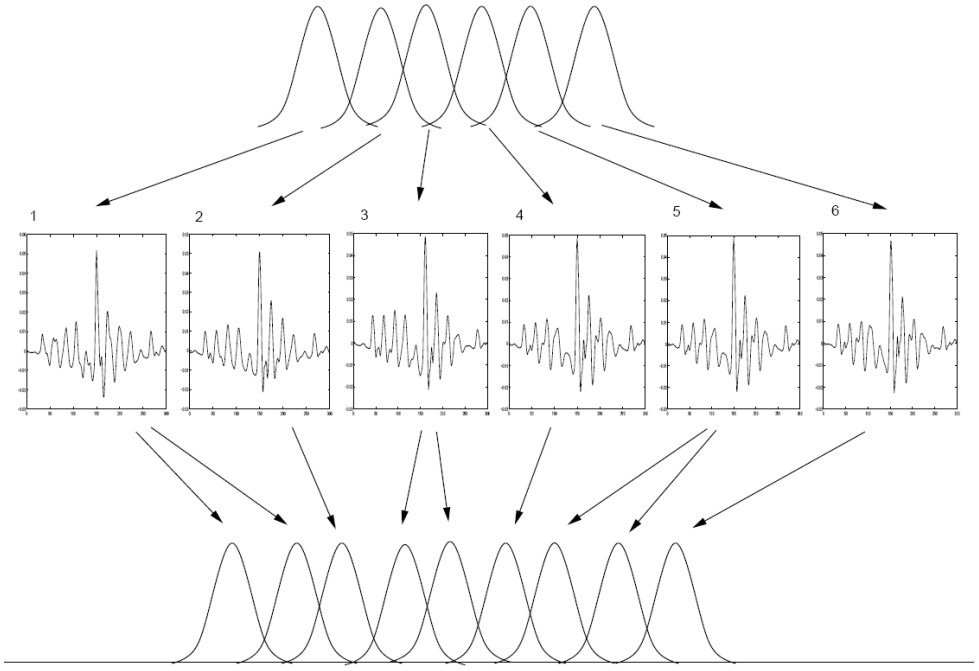

Figura 5.3: Escalado de la duración

Ejercicio 6 *La función "tdpsola.m" permite modificar la frecuencia fundamental y duración de una señal de voz por unos factores de escalado constantes. Sobre el fichero "iago800_0001.wav" realiza distintas manipulaciones de únicamente frecuencia fundamental, únicamente duración y conjuntamente frecuencia fundamental y duración. Escucha los resultados y compara segmentos de la forma de onda resultante con los correspondientes en la forma de onda original.*

Ejercicio 7 *El programa Praat permite realizar el marcado de pitch y manipulación con el algoritmo TD-PSOLA de forma sencilla y flexible. Siguiendo las instrucciones de tu profesor generarás entre otras señales:*

- *Señal de voz con el contorno entonativo estilizado (aproximado).*
- *Señal de voz con frecuencia fundamental constante.*
- *Señales con contorno entonativo y duración modificados localmente.*

Ejercicio 8 *Suponiendo que dispones del módulo lingüístico-prosódico de un conversor texto-voz que te proporciona la secuencia de sonidos a generar y la información fonética asociada,*

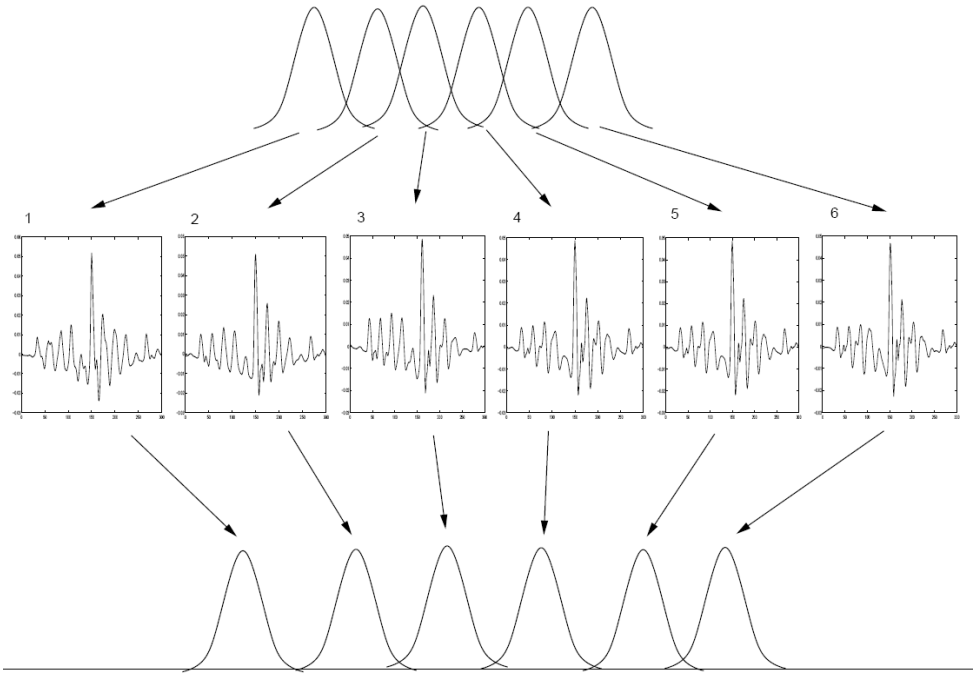

Figura 5.4: Escalado de la frecuencia fundamental

explica los pasos que seguirías para construir un conversor texto-voz por concatenación de unidades. Considera como unidades difonemas, teniendo una única realización de cada uno de ellos, y utiliza la técnica TD-PSOLA.

Bibliografía

[Taylor, 2008] Taylor, P. (2008). *Text-to-speech Synthesis*. Cambridge University Press.