Síntesis de voz utilizando difonemas: Uniones entre vocales.

Roger Guaus i Termens, Jaume Oliver i Lafont, Francesc Gudayol i Pontabella, Josep Martí i Roca

Secció de Tecnologies de la Parla
Departament d'Acústica
Escola d'Enginyeria La Salle
Universitat Ramon Llull

Introducción

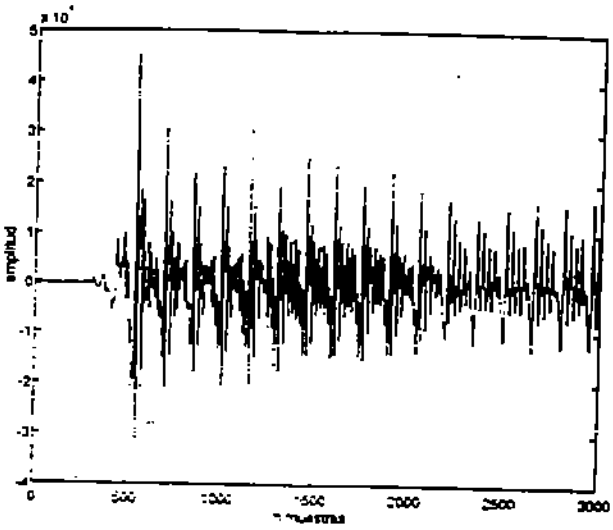
Los métodos estudiados recientemente para realizar síntesis de habla utilizan segmentos de voz pregrabados en una base de datos, que se concatenan de forma adecuada en el momento de la síntesis. Las unidades más utilizadas hasta el momento han sido fonemas, sílabas, semisílabas y difonemas. El hecho de escoger un tipo de unidades u otro viene dado por un compromiso. En general una lengua romànica suele tener unos 40 alófonos, así que si utilizáramos los alófonos como unidades básicas de síntesis obtendríamos una base de datos de unidades de pequeñas dimensiones y fácilmente manejable. Sin embargo, la calidad de la voz obtenida en este caso no sería la deseada. Podemos ver intuitivamente que cuan mayor sean las unidades, mayor será la calidad de la voz generada. Pero si utilizáramos palabras o frases pregrabadas para sintetizar cualquier tipo de mensaje, las dimensiones de la base de datos serían demasiado grandes y poco manejables. De esta manera se llega a un compromiso de calidad - memoria que nos lleva a trabajar con difonemas. Los difonemas son unidades de voz que contienen dos alófonos (p.e. combinaciones vocal + vocal, voval+consonante, etc.). En el momento de la síntesis se utiliza la parte de los difonemas comprendida entre el centro del primer alófono y el centro del segundo. De esta manera tenemos las transiciones entre fonemas pregrabadas y la unión entre difonemas se realiza en las partes estables del alófono. El hecho de concatenar los difonemas por sus partes estables facilita el procesado que se realiza para la unión, así como la modificación de la duración de cada alófono. En algunos casos particulares donde los alófonos carecen de parte estable se utilizan trifonemas (p.e. en las combinaciones oclusiva + *r* + vocal, la *r* es un alófono inestable que se considera como si fuera una transición entre la oclusiva y la vocal). Así pues, las unidades de síntesis pueden comprender indistintamente difonemas o trifonemas.

Problema de la unión de difonemas

El problema principal de este método es el efecto de la coarticulación que presentan los alófonos. Las 1200 unidades que necesita el catalán para sintetizar cualquier tipo de mensaje se han obtenido de una grabación de 1200 palabras y frases. En el momento de la síntesis concatenamos dos difonemas por las partes estables de dos alófonos iguales pero que se han grabado en contextos distintos de coarticulación. Este hecho provoca una discontinuidad en la evolución temporal del espectro del alófono que repercute claramente en la calidad de la voz generada.

Veamos un ejemplo:

Palabra	Transcripción	Difonema
pegot	[p@GOt]	[GO] + [Ol]
solta	[sOlt@]	

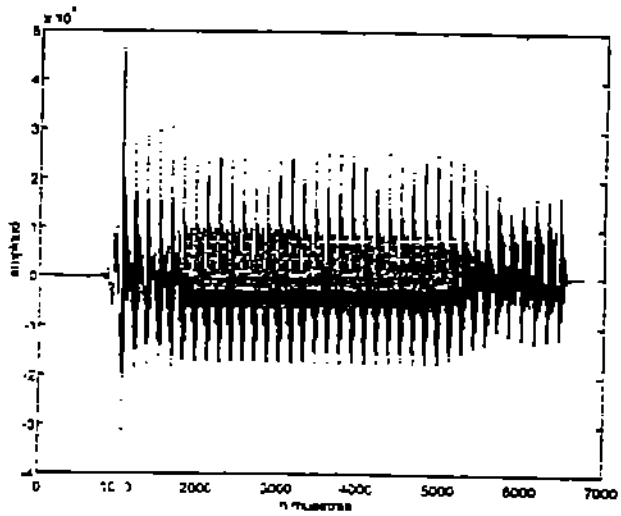


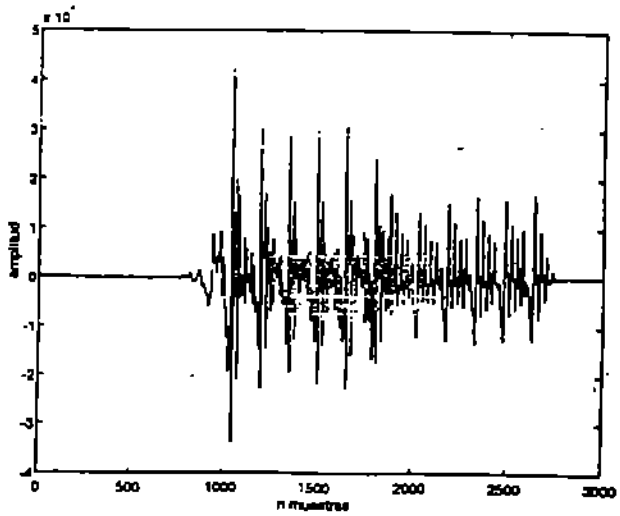
En la figura 1a vemos las dos secuencias temporales correspondientes al los difonemas [GO] y [OI] para generar la palabra [GOI]. La figura 1b muestra la evolución temporal de los formantes de los difonemas y la discontinuidad debida a los distintos contextos de articulación en el momento de la grabación.

Problema de la modificación de la duración

Otro problema que afecta claramente la calidad de la voz es la modificación del parámetro prosódico de la duración de los alófonos. Los difonemas almacenados en la base de datos tienen una duración determinada por el momento de la grabación. Así que si deseamos generar alófonos de mayor o menor duración, habrá que procesar de alguna manera la señal de voz para conseguirlo. Los métodos de síntesis de voz trabajan con tramas de señal de 15 a 40 milisegundos aproximadamente. En nuestro caso, utilizando el método de síntesis PSOLA (1), las tramas de señal equivalen a dos periodos de la señal (análisis en banda ancha) multiplicadas por una ventana Hanning de la misma longitud. De modo que cuando se desea aumentar la duración de un alófono se repite la última trama del primer alófono y la primera del segundo. En principio, al tratarse de la parte estable del alófono parece que la solución es correcta, sin embargo aparecen efectos indeseados que merman la calidad del habla generada. Al repetir varias veces la misma trama idéntica se genera una señal perfectamente periódica que produce una sensación extremadamente sintética. Además, el hecho de tener una evolución espectral constante produce un efecto de discontinuidad en la unión de los dos difonemas mucho más acentuada que en un caso de duración no modificada.

Por otra parte, cuando intentamos disminuir en exceso la duración de la señal, eliminando las últimas tramas del primer alófono y las primeras del segundo, puede ocurrir que eliminemos por completo la parte estable del alófono, con lo cual obtendríamos una señal compuesta únicamente por transiciones.





En la figura vemos los casos de aumentar y disminuir la duración de un alófono.

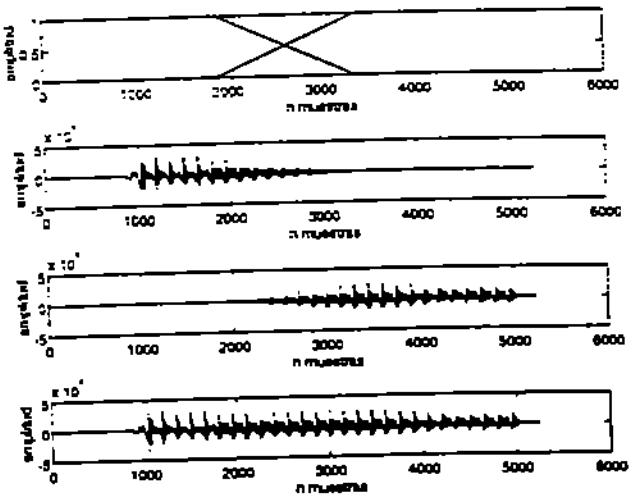
Soluciones al problema de la unión de los difonemas.

El efecto de la discontinuidad se percibe de manera más intensa en el caso de las vocales debido, en general, a su mayor estabilidad y duración. Por lo tanto en adelante estudiaremos distintas soluciones para el caso concreto de una unión de dos difonemas del tipo [consonante - vocal] + [vocal - consonante] y el punto de unión será siempre una vocal.

Disponemos de dos métodos para mejorar las transiciones entre difonemas.

Método de la mezcla de tramas:

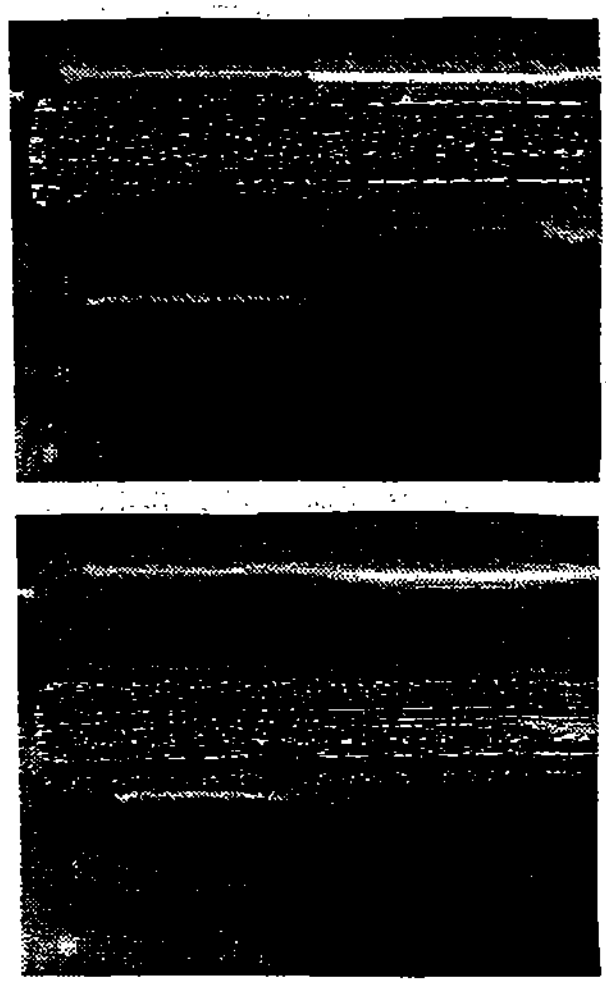
Mantener el punto de unión en el centro del alófono vocálico y realizar una combinación lineal de las últimas tramas de la primera vocal con las primeras tramas de la segunda. De esta manera se suaviza la transición entre los dos difonemas y el efecto de discontinuidad disminuye. La combinación se realiza con dos factores de ponderación de variación lineal en forma de rampas complementarias de modo que la suma de las dos resulta la unidad.



En estas figuras vemos las funciones que realizan la combinación lineal y su efecto final.

Este proceso temporal repercute en el dominio de las frecuencias de manera que los formantes del primer alófono varían progresivamente hasta el segundo.

La mejora que introduce este método es más apreciable cuanto mayor sea la duración del fonema vocálico. Si evaluamos solo los dos difonemas generados, en situaciones donde las duraciones de las vocales son cortas (de 60 a 80 ms), los efectos resultantes aunque justificados matemáticamente suelen ser poco perceptibles debido a la poca estabilidad en la evolución espectral de la señal. Sin embargo este método introduce una mejora notable en la calidad de la voz de una frase sintetizada.



La primera figura muestra el sonograma de una unión sin interpolación. La segunda con una interpolación de 8 tramas.

Método de unión en un extremo:

En algunos casos se puede utilizar la totalidad de la vocal sin tener que crearla a partir de dos segmentos distintos, y unir el extremo de la vocal directamente con su consonante adyacente.

Pero para realizar este tipo de unión se tienen que cumplir las dos condiciones siguientes. Imaginemos que deseamos realizar la unión por la derecha de los difonemas [la] y [af] para obtener [laf]. Primera condición: En el difonema [la], la vocal grabada de origen debe estar mínimamente coarticulada por su consonante siguiente (p.e. sería válida la

grabación [lap], pero no [lal]). Segunda condición: La consonante del segundo difonema a sintetizar debe tener el mismo punto de articulación que la consonante grabada en la derecha (p.e. si se cumpliera la primera condición, sería válido sintetizar [laf] pero no [laj]). Así que uniendo por la derecha obtendríamos [la] [f].

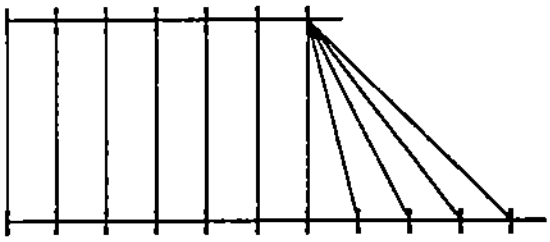
Palabra	Transcripción	Síntesis
elàstic	[lapse]	[la] + [f]
Olaf	[ulaf]	

En caso de unir por la izquierda se aplicarían las mismas condiciones a la primera consonante.

Modificación de la duración de la vocal.

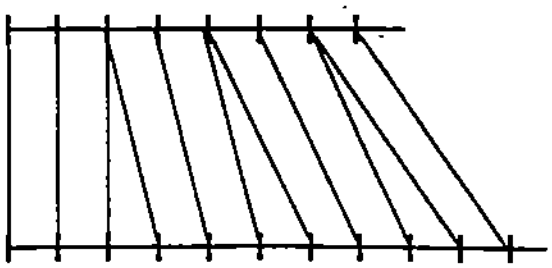
Aumento de la duración

Hemos visto que repetir las tramas que se encuentran en los extremos de la vocal con intención de aumentar la duración del fonema produce efectos perceptibles no deseados. Una solución válida es la aportada por Moulines (1,2) y Charpentier (1) en su trabajo sobre el sistema de síntesis PSOLA donde proponen, en lugar de repetir únicamente las tramas de los extremos para aumentar la duración del alófono (muchas repeticiones), repetir tramas intercaladas a lo largo de toda la parte estable del alófono.



La primera figura muestra en el eje horizontal la repetición de tramas según el método inicial.

En caso de aumentar excesivamente la duración, este método provoca una evolución temporal del espectro discontinua debida a la repetición sucesiva de tramas que puede ser claramente perceptible. Proponemos realizar una interpolación de tramas intercaladas a partir de combinaciones lineales de sus adyacentes de manera que ninguna de las tramas de la vocal sea igual a otra y la evolución de los formantes sea perfectamente continua.



La figura muestra el método para aumentar la duración propuesto por Moulines y Charpentier.

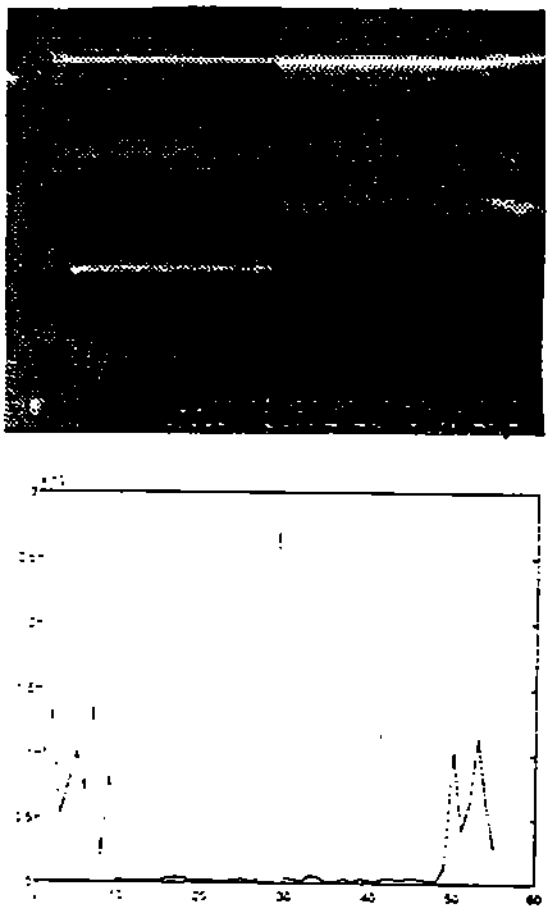
Disminución de la duración

En este caso se pueden aplicar dos de los conceptos ya vistos: Eliminar tramas de los extremos de la vocal o bien eliminarlos intercaladamente a lo largo de toda su parte estable. El mayor problema se presenta ante la intención de reducir excesivamente la duración de la vocal ya que se corre el riesgo de eliminar por completo la parte estable del fonema y obtener una habla generada únicamente a partir de transiciones entre consonantes y vocales. Este hecho depende claramente del locutor que ha sido grabado y su forma de articular. Generalmente se desea que el locutor pronuncie con ritmo moderado con el fin de obtener segmentos de voz con partes estables lo suficientemente largas para poder procesar la señal sin problemas en el momento de la síntesis. Sin embargo se ralentizan también las transiciones entre unos alófonos y otros, con lo cual estas transiciones suelen ser bastante lentas, efecto que produce, en el caso de disminuir excesivamente la duración del alófono, que las transiciones sean de mayor duración que las partes estables. Aunque hay transiciones que son invariables con el ritmo del habla (r, f + r + a, etc.), se debe considerar también la supresión de algunas tramas de los puntos de unión entre fonemas consonánticos y vocálicos. En general este efecto se produce en las vocales que se encuentran fuertemente coarticuladas con sus consonantes adyacentes.

Conclusiones

En el momento de implementar el sistema hay que contemplar la solución de los dos problemas que se presentan: la unión de los difonemas y la modificación de la duración. La técnica de las uniones por combinación lineal se puede utilizar con cualquiera de las

técnicas de la modificación de la duración. En este caso es necesario determinar la longitud del segmento a mezclar en función de la duración de la parte estable. Por otra parte, en las uniones realizadas en los extremos (derecha o izquierda) no tiene sentido realizar combinación lineal de las tramas debido a la poca semejanza que tienen las tramas de la transición. Una manera de medir la continuidad de los formantes en los puntos de unión es aplicar una medida de distancia espectral entre un espectro en un instante de tiempo y otro espectro inmediatamente adyacente. Este proceso nos dará una curva que mostrará la discontinuidad del espectro a lo largo de la evolución temporal.



En la figura inferior podemos ver la medida de la distancia espectral que existe en el sonograma de la figura. Se distingue claramente la discontinuidad espectral.

Una vez estudiados los distintos problemas y soluciones nos planteamos como línea de futuro implementar un sistema capaz de decidir que tipo de solución se debe aplicar en cada unión de difonemas que se presente. Este sistema deberá funcionar *on-line*, es decir, a medida que se sintetiza el habla se va decidiendo para cada difonema cual es el sistema óptimo para la unión.

Referencias

(1) Moulines, E., Charpentier, F. *Pitch-Sinchronous waveform processing techniques for text-to-speech synthesis using diphones.* Speech Communication 9 (1990) 453-467.

(2) Moulines, E., Laroche, J. *Non-parametric techniques for pitch-scale and time-scale modification of the speech.* Speech Communication 16 (1995) 175-205.

(3) Dutoit, T., Leich, H. *Inproving the TD-PSOLA Text-To-Speech synthesizer with a specialy designed MBE Re-Synthesis of the Segments Database.* Signal processing VI: Theories and Applications (1992).

(4) Dutoit, T., Leich, *H. MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database.* Speech Communication 13 (1993) 435-440.

(5) Valbret, H. Moulines, E. Tubach, J.P. *Voice transformation using PSOLA technique.* Speech Communication 11 (1992) 175-187.

(6) O'Shaughnessy, D., Barbeau, L., Bernardi, D., Archambault, D. *Diphone speech synthesis.* Speech Communication 7 (1988) 55-65.

(7) Klatt, D. *Review of text-to-speech conversion for English.* Journal Acoustics Society of America 82 (1987) 737 - 793.