ACÚSTICA

Análise de Voz

SÍNTESIS DE VOZ POR CONCATENACIÓN DE UNIDADES: MEJORAS EN LA CALIDAD SEGMENTAL.

Roger Guaus i Térmens, Jaume Oliver i Lafont,
Helena Moure González, Ignasi Iriondo Sanz, Josep Martí i Roca

Secció de Tecnologies de la Parla, Departament d'Acústica
Escola d'Enginyeria La Salle, Universitat Ramon Llull
Pg. Bonanova 8, 08022 Barcelona, Catalunya
Tel: +34 932902405 / Fax: +34 932902416
E-mail: rogerg@salleURL.edu

SUMMARY

Recent approachments in speech synthesis work with recorded segments of speech that are coupled during the synthesis process. Coupling diphones is not an easy work, so here we describe some methods to achieve it in order to improve the segmental quality of the synthesized speech.

INTRODUCCIÓN

Las últimas tendencias al realizar síntesis de voz se basan en la concatenación de unidades pregrabadas [5], [6], [7] y [8]. Según la unidad escogida tendremos un compromiso entre la calidad segmental del habla generada y la magnitud de la base de datos que contiene la voz grabada. Aunque han surgido recientemente nuevas ideas sobre selección de unidades que tratan de mejorar la calidad segmental del habla, este trabajo está enfocado para obtener el máximo rendimiento de los sistemas de síntesis basados en concatenación de difonemas y trifonemas [1], [2] y [3].

UNIÓN DE DIFONEMAS Y TRIFONEMAS

En el momento de la síntesis concatenamos dos difonemas por las partes estables de dos alófonos iguales, pero grabados 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 [4].

El efecto de la discontinuidad se percibe de manera más intensa en el caso de las vocales debido, generalmente, a su mayor estabilidad y duración. Por lo tanto, estudiaremos en adelante 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.

Sea $[C_{a1}V_1]$-$[V_2C_{p2}]$ la unión entre dos difonemas, donde V_1 y V_2 representan la misma vocal de dos difonemas distintos, grabados en momentos distintos, y C_{a1} y C_{p2} las consonantes anterior del primer difonema y posterior del segundo, respectivamente. Por lo tanto la unidad sintetizada se unirá por algún punto intermedio de V_1 y V_2. Definamos también las unidades en su entorno de grabación como $[C_{a1}V_1\,X_{p1}]$ y $[X_{a2}V_2C_{p2}]$, donde

X_{p1} y X_{a2} son los alófonos posterior del primer difonema y anterior del segundo que no pertenecen a la unidad, pero que han influido claramente en la articulación de la vocal.

Unión por el extremo

En algunos casos se puede utilizar únicamente la totalidad de una de las dos vocales, sin tener que crear la unión de los dos difonemas a partir de dos segmentos distintos y unir el extremo de la vocal directamente con la consonante del otro difonema, asegurando de este modo que no habrá transición ninguna en la vocal. Este tipo de uniones pueden realizarse por la derecha o por la izquierda, pero sólo cuando las condiciones articulatorias son favorables.

Para realizar la unión por la izquierda o por la derecha hay que evaluar la distancia espectral entre los alófonos C_{a1} y X_{a2} para la unión izquierda y X_{p1} y C_{p2} para la derecha, formulándolas respectivamente como:

$$D_i = \text{dist}\{C_{a1}, X_{a2}\}, \quad D_d = \text{dist}\{X_{p1}, C_{p2}\}$$

Que en la unión por la izquierda el valor de D_i sea pequeño significa que la distancia, en la parte izquierda, entre la evolución del espectro de la vocal grabada y la evolución del espectro de la vocal deseada será pequeña y, por consiguiente, se podrá utilizar la totalidad de la vocal V_2, realizando una unión en el extremo izquierda de la forma $[C_{a1}V_2C_{p2}]$ y rechazando de esta manera la utilización de V_1. Para unir por la derecha se utilizará el mismo criterio, de manera exista una distancia pequeña D_d entre X_{p1} y C_{p2}. En la unidad sintetizada se obtendría $[C_{a1}V_1C_{p2}]$.

En estos casos, la unión se realiza en un punto de inestabilidad en la evolución temporal del espectro entre la consonante y la vocal, mientras que el espectro de la vocal tendrá una evolución continua sin transiciones que mermen la calidad deseada.

La medida de distancia se puede basar en distintos criterios. Por una parte se pueden definir distancias de manera simple, en función del punto de articulación de cada consonante. En este caso, tomando en cuenta las posiciones geométricas de los articuladores, definiendo las posiciones de bilabial, labio-dental, dental, alveolar, etc. y definiendo una distancia en función de la separación entre estos puntos. Otra solución es medir la distancia espectral objetiva entre las dos consonantes a partir, por ejemplo, de los coeficientes Cepstrum.

Una vez obtenido el valor de la distancia mínima entre las dos uniones posibles (izquierda o derecha) tenemos que optar por la mejor de las dos:

$$D_t = \min\{D_i, D_d\}$$

Imaginemos, por ejemplo, que deseamos unir los difonemas [sa] y [ak] provenientes de las grabaciones $[sa_1l]$ y $[za_2k]$, respectivamente. Si evaluamos la distancia D_i=dist{s,z} y D_d=dist{l,k}, obtendremos que $D_i<D_d$, por lo que será preferible realizar la unión por el extremo izquierdo $[sa_2k]$.

Si la distancia obtenida queda por debajo del umbral preestablecido empíricamente, se realizará la unión por dicho punto. Si D_t supera el umbral deberemos buscar otro sistema de unión como el que describimos a continuación.

Mezcla de tramas

En caso de no poder realizar la unión por alguno de los dos extremos, se busca un punto de unión donde hacer la transición de la vocal V1 a V2. Típicamente se marcaba un punto de la vocal que se consideraba más estable y se realizaba la unión en aquel punto. Pero este método no aseguraba que en la transición la evolución del espectro fuera adecuada. Por este motivo, proponemos realizar la unión en un punto óptimo calculado a partir de las distancias entre los coeficientes Cepstrum de las tramas de cada una de las vocales.

Imaginemos que las vocales V1 y V2 están formadas por M y N tramas de señal, respectivamente. De esta manera podemos calcular las distancias Dij donde i=1, 2,.., M y j=1, 2,.., N, y crearemos una matriz de MxN elementos.

$$\underline{\underline{D}} = \begin{pmatrix} D_{11} & D_{21} & . & D_{M1} \\ D_{12} & . & . & . \\ . & . & . & . \\ D_{1N} & . & . & D_{MN} \end{pmatrix}$$

El punto de unión estará situado donde se obtenga un mínimo de distancia espectral:

$$(i, j) = \arg min(\underline{\underline{D}})$$

De todos modos esta operación, aunque matemáticamente dé un valor correcto, puede proporcionar un punto de unión con poco sentido físico, según los valores de i y j. Este hecho implica plantear restricciones en la búsqueda del punto óptimo y, por lo tanto, nos centraremos en el entorno de $i=j$ para $i=M/2$ aproximadamente; es decir, en las partes estables de cada alófono.

Una vez encontrado el punto de unión (i,j), se efectua la unión propiamente dicha. Se realiza 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 en cada momento resulta la unidad. 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 sólo los dos difonemas generados, en situaciones donde las duraciones de las vocales son cortas, 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, desde el punto de vista global, en la calidad de la voz de una frase sintetizada.

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, deberemos 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.

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 que obtendríamos una señal compuesta únicamente por transiciones.

Normalmente, los métodos de repetir o eliminar tramas de señal intercaladamente para aumentar o disminuir la duración de un alófono respectivamente, provocan un efecto de poca naturalidad en la señal generada. Para minimizar este tipo de efecto, hemos desarrollado un método de conversión de número de tramas de señal.

Para convertir una señal de N tramas a una de M, donde N puede ser mayor, igual o menor que M, podemos crear una matriz de dimensiones N x M que contenga unos coeficientes de transformación a_{ij} que determinen la proporción de la trama x_i que se debe aplicar sobre la trama y_j.

Así pues la trama y_j será una combinación lineal de las N tramas $x_1..x_N$. De esta manera no habrá repetición de tramas y, debido a la mejor evolución temporal del espectro, obtendremos mejor calidad que en los casos descritos anteriormente. Para que no haya problemas de energía en las tramas generadas, es necesario que la suma de los coeficientes que forman una trama y_j sea constante. Así pues, una premisa para definir los coeficientes es

$$\sum_{i=1}^{N} a_{ij} = 1 \text{ para } j = 1..M$$

y de esta manera quedan los coeficientes definidos en el dominio $a_{ij} \in [0,1]$.

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_M \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & . & . & a_{N1} \\ a_{12} & a_{22} & . & . & a_{N2} \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{1M} & a_{2M} & . & . & a_{NM} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_N \end{pmatrix}$$

Pero esta condición no es suficiente para definir a_{ij}, ya que se desea que la evolución temporal de los espectros de las tramas varíe de acuerdo con la evolución de la señal original. La diferencia de índices *i-j* da una idea de la influencia de la trama x_i sobre la trama y_j.

Así pues, podemos concluir que los valores de a_{ij} con una diferencia de índices grande, tienen que ser muy pequeños o nulos, debido a que la influencia de x_i sobre y_j debe ser pequeña. Por otra parte, si la diferencia de índices es un valor pequeño, el valor del coeficiente a_{ij} será grande. Por lo tanto, para definir los coeficientes a_{ij} se utilizan muestras de funciones del tipo creciente - decreciente, situando el máximo de la función sobre la trama de máxima influencia.

Este sistema, a partir de la multiplicación matricial, presenta muy buenos resultados ya que se asegura una evolución continua del espectro en todo momento, creando M tramas nuevas a partir de una combinación lineal de las N tramas originales, evitando los inconvenientes de los sistemas tradicionales de simple yuxtaposición.

CONCLUSIONES

En el momento de implementar el sistema debemos contemplar la solución de los dos problemas presentados, funcionando de forma simultanea: 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.

Otra forma de hallar el punto de unión entre dos alófonos es considerar la duración final deseada del alófono y buscar el punto de unión considerando la duración de los alófonos grabados; sin embargo este sistema no ha sido comentado debido a los pobres resultados obtenidos.

Por otra parte sería interesante extender la utilización del método de mezcla de tramas, no sólo a las uniones en las partes estables de los alófonos, sino también cuando se realizan uniones por los extremos, y de esta manera suavizar la evolución temporal del espectro en los casos que sean necesarios.

Una vez estudiados los distintos problemas y soluciones, nos planteamos como línea de futuro aplicar estas técnicas en los recientes sistemas de selección de unidades.

REFERENCIAS

1. Conkie A.D., Isard, S. *Optimal Coupling of Diphones.* Progress in Speech Synthesis. 1997 Springer-Verlag New York, Inc.
2. 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).
3. 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.
4. Klatt, D. *Review of text-to-speech conversion for English.* Journal Acoustics Society of America 82 (1987) 737 - 793.
5. Moulines, E., Charpentier, F. *Pitch-Sinchronous waveform processing techniques for text-to-speech synthesis using diphones.* Speech Communication 9 (1990) 453-467.
6. Moulines, E., Laroche, J. *Non-parametric techniques for pitch-scale and time-scale modification of the speech.* Speech Communication 16 (1995) 175-205.
7. Portele, T., Höfer, F., Hess, W.J. *A mixed inventory structure for german concatenative synthesis.* Progress in Speech Synthesis. 1997. Springer-Verlag New York, Inc.
8. O'Shaughnessy, D., Barbeau, L., Bernardi, D., Archambault, D. *Diphone speech synthesis.* Speech Communication 7 (1988) 55-65.