Procesamiento del Lenguaje Natural, num. 29 (2002), pp.159-163 recibido 25-05-2002; aceptado 02-07-2002

Un modelo híbrido orientado a la síntesis multimodal del habla

Ignasi Iriondo Sanz, Francesc Alías Pujol, Javier Melenchón Maldonado
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
Pg. Bonanova 8 08022 Barcelona
{iriondo, falias, jmelen}@salleURL.edu

Resumen: En este artículo se presenta un sistema de conversión texto-habla de alta calidad utilizando voz segmentada en difonemas y trifonemas. El sistema de síntesis implementado se basa en un modelo híbrido que combina aspectos de un modelo "armónico + ruido", con el que se descompone la señal de voz original en dos componentes, y aspectos del TD-PSOLA. Los procesos de análisis y síntesis se realizan síncronamente con el *pitch*, de forma que se pueden conseguir modificaciones prosódicas con un alto grado de naturalidad en el habla generada gracias a la representación paramétrica de la señal de voz. Este sistema resulta una buena solución para la síntesis del habla emocionada, que requiere grandes variaciones de la prosodia. El objetivo final de este proyecto consiste en implementar este modelo híbrido de síntesis en un sistema de síntesis audiovisual del habla, capaz de generar síncronamente voz y animación facial para simular expresiones emocionales.
Palabras clave: Síntesis concatenativa del habla, modelo "armónico + ruido", prosodia, síntesis audiovisual del habla, modelos de la expresión emocional

Abstract: In this paper we present a high-quality text-to-speech system using diphones and triphones. The implemented synthesis system is based on a hybrid model that combines a harmonic plus noise decomposition technique with some features of TD-PSOLA. The analysis and the synthesis processes are pitch-synchronous, so prosodic modifications can be generated achieving a more natural-sounding of synthetic speech. This parametric representation of speech outperforms other techniques for concatenative synthesis (e.g., TD-PSOLA) in intelligibility and naturalness, so it is a good solution for emotional speech synthesis, which requires high-quality prosody modifications. The final goal of this project is to integrate this hibrid speech synthesis method in a text-to-audiovisual speech system that can generate synchronously speech and facial animation to emulate emotional expressions.
Keywords: Concatenative speech synthesis, harmonic plus noise model, prosody, text-to-audiovisual speech, emotional expression modeling

1. Introducción

La síntesis multimodal del habla, o también llamada síntesis audiovisual, consiste en la generación automática de voz y la correspondiente animación facial a partir de un texto cualquiera. Su principal función es la de servir como interfaz de sistemas de diálogo hombre-máquina. El hecho de mostrar la cara de una persona moviendo la boca de forma sincronizada con el habla generada, favorece la comprensión del mensaje, especialmente en situaciones con deficientes condiciones acústicas. Además, se puede mejorar la naturalidad del sistema añadiendo expresiones faciales que conlleven una significación del estado emocional del hablante virtual. Estas expresiones faciales se deben acompañar de una voz con la misma carga emotiva para conseguir una buena credibilidad por parte de los usuarios. El proceso de síntesis del habla producida con diversas emociones conlleva diferentes niveles de actuación. En primer lugar, se requiere una adecuación del texto que incluya un vocabulario propio del estado emocional que se desea transmitir. Esta capacidad tiene que estar implementada en el sistema de diálogo, que será el encargado de generar el texto que se debe sintetizar. En segundo lugar, se requiere de un modelo acústico de la expresión emocional para el idioma en cuestión, que permita la generación automática de la prosodia adecuada. Este modelo acústico tiene que establecer básicamente:

ISSN 1135-5948

la curva de entonación, la duración de los sonidos, la fragmentación del discurso, la duración de la pausas y las variaciones de la energía (Rodríguez et al., 1999). Por último, la inclusión de expresión emocional en el habla comporta un aumento en la variabilidad de los parámetros prosódicos (*pitch*, duración y energía), haciéndose necesario un sistema de síntesis capaz de generar dichas modificaciones prosódicas minimizando en la medida que sea posible la distorsión del habla generada.

Los métodos más utilizados en síntesis concatenativa aplicada a conversión texto-habla (CTH) se basan las técnicas TD-PSOLA (Moulines y Charpentier, 1990) y MBROLA (Dutoit y Leich, 1996). TD-PSOLA lleva a cabo un procesado síncrono con el *pitch* de la señal de voz, consiguiendo una buena calidad en el habla sintetizada. Sin embargo, esta técnica presenta una serie de limitaciones inherentes a su estructura no paramétrica, fundamentalmente la discontinuidad espectral y de frecuencia fundamental que presenta en los puntos de concatenación (Syrdal et al., 1998). Además, no se consiguen resultados satisfactorios cuando las modificaciones prosódicas son importantes. MBROLA intenta solucionar dichas limitaciones resintetizando las unidades sonoras con *pitch* y fase constantes. Este proceso artificial causa un zumbido en el habla sintetizada que no permite mejorar de forma sustancial los resultados obtenidos con TD-PSOLA (Syrdal et al., 1998). El método HNM (*Harmonic plus Noise Model*), propuesto por Stylianou (1996b), consigue modificaciones prosódicas de alta calidad basándose en una representación parámetrica de la señal de voz que, mediante un análisis síncrono con el *pitch*, la descompone en una parte armónica y otra de ruido. La aplicación de esta técnica en un sistema de CTH se presenta en (Stylianou, Dutoit, y Schroeter, 1997) y en (Stylianou, 2001). El primero se basa en una concatenación de difonemas y el segundo utiliza un método de selección de unidades que le permite minimizar la modificación de la prosodia original. Este cambio en la estrategia de síntesis se debe a la evolución que ha experimentado los sistemas de CTH en los últimos años.

El presente artículo describe el módulo de síntesis de un sistema CTH en catalán basado en síntesis concatenativa mediante un modelo híbrido entre la técnica TD-PSOLA y el método HNM, con el objetivo de conseguir un control del *pitch* mucho más eficiente que en nuestro sistema anterior basado únicamente en TD-PSOLA. La primera parte del artículo se dedica a la descripción del modelo híbrido de parametrización del habla. A continuación, se describe el proceso de síntesis concatenativa basada en difonemas y trifonemas para llevar a cabo la CTH. Y por último, se exponen los resultados y conclusiones que se han obtenido al aplicar dicho método en un sistema de síntesis audiovisual del habla capaz de generar expresiones emocionales tanto acústica como visualmente.

2. *El modelo híbrido*

2.1. Descripción del modelo

El modelo híbrido presentado se basa en un proceso de análisis y síntesis del habla síncrono con el *pitch*, similar a TD-PSOLA, pero con la diferencia de que se descompone la señal de voz en una parte armónica y otra de ruido siguiendo la filosofía de HNM. Este sistema constituye un primer paso en el camino de mejora de nuestro sistema de CTH orientado a conseguir una mayor naturalidad en la síntesis del habla emocionada.

Un modelo "armónico + ruido" parte de la suposición que la señal de voz se compone de una parte armónica $h(t)$ y una parte de ruido $n(t)$. La parte armónica modela la componente casi periódica de la señal de voz, mientras que la parte de ruido modela sus componentes no periódicas, tales como el ruido de fricción y los sonidos no sonoros.

Desde un punto de vista frecuencial, esta descomposición supone una división del espectro en dos bandas, separadas por la máxima frecuencia sonora $F_m(t)$ que será un parámetro variante, tal como se puede observar en el ejemplo de la figura 1.

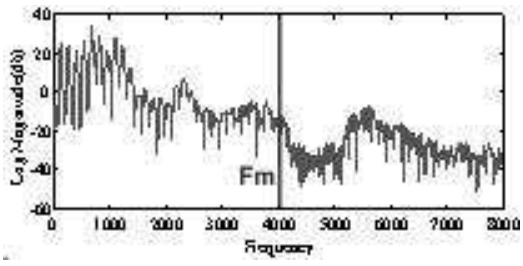

Figura 1: Espectro de la señal de voz

La banda inferior del espectro se representa únicamente por sinusoides armónica-

mente relacionadas entre sí (f_o es la frecuencia fundamental) como indica la fórmula 1. La amplitud $A_k(t)$ y fase $\phi_k(t)$ del armónico k-ésimo serán variantes en el tiempo.

$$h(t) = \sum_{k=1}^{K(t)} A_k\,(t)\cos(2\pi k f_o t + \phi_k(t)) \quad (1)$$

A su vez, la banda superior, que contiene mayoritariamente la componente de ruido, se puede asociar a un modelo AR que permitirá sintetizar esta parte mediante un filtro todo polos variante en el tiempo, como en el modelo HNM_1 presentado por Stylianou (2001). Otra opción para modelar la parte de ruido consiste en no parametrizarla y generarla directamente como la diferencia entre la señal original y la componente armónica.

$$n(t) = s(t) - h(t) \quad (2)$$

La ventaja de esta aproximación es que también se tiene en cuenta la parte de ruido presente en la banda inferior del espectro, aunque al no parametrizar la señal, se pierde la compresión de información obtenida con el modelo AR.

2.2. Análisis de la señal de voz

El análisis de la señal de voz se lleva a cabo de forma síncrona con el *pitch*, es decir, se requerirá de un sistema automático de asignación de marcas del *pitch* (Alías y Iriondo, 2001). La señal de voz se analizará mediante tramas centradas en estas marcas (t_i) y de una duración de dos periodos de *pitch*. El sistema utilizado se caracteriza por marcar tanto las zonas sonoras como las sordas o los silencios. De esta forma se evita el proceso de distinguir previamente la sonoridad de la señal de voz. En las zonas periódicas, las marcas se sitúan sobre el pico de amplitud máxima y en las zonas no periódicas se distribuyen formando una transición desde el valor de *pitch* precedente al valor posterior.

El siguiente paso es el cálculo de la frecuencia máxima sonora F_m, que determinará el número de armónicos que formarán parte del modelo. Stylianou (1996a) propone un método de cálculo de F_m a partir de un análisis en el dominio frecuencial mediante el cual se decide si los picos del espectro forman parte de la componente sonora o no. El último de estos picos sonoros se corresponde a F_m. Con la finalidad de simplificar el proceso de análisis, y teniendo en cuenta que la parte de ruido cubre todo el espectro, se ha decidido mantener constante el número de armónicos, con lo que el cálculo de F_m es directamente el producto de la frecuencia fundamental por el número de armónicos.

En la estimación de las amplitudes y fases de la parte armónica se asume que ambas son constantes para toda la trama de análisis, al igual que el periodo de *pitch*. Para facilitar el cálculo de la parte armónica se utiliza la fórmula 3 en la cual la amplitud A_k es compleja y por tanto incorpora información del módulo y la fase.

$$h(t) = \sum_{k=-L}^{L} A_k\left(t_a^i\right) e^{j2\pi k f_o(t_a^i)(t-t_a^i)} \quad (3)$$

donde L es el número de armónicos, A_k es la amplitud compleja del armónico k-ésimo y cumple $A_{-k} = A_k^*$. El cálculo de dichas amplitudes complejas se realiza mediante un método ponderado de mínimos cuadrados que minimiza la fórmula 4 respecto A_k.

$$\varepsilon = \sum_{t=t_a^i-N}^{t_a^i+N} w^2\,(t)\;(s(t) - h(t))^2 \quad (4)$$

donde $w(t)$ es la función de ponderación (en nuestro caso, una ventana *hanning*), y N es el entero más próximo al periodo de *pitch* local. Es importante remarcar que la ventana de análisis está centrada en el instante de tiempo t_a^i y la longitud de la ventana es $M = 2N+1$.

Finalmente, la componente de ruido se calcula como la diferencia entre la señal de voz y la parte armónica.

2.3. Síntesis de la señal de voz

La síntesis de la señal de voz también se lleva a cabo de forma síncrona con el *pitch* mediante un proceso complementario al de análisis. En primer lugar se reconstruye la componente armónica correspondiente a cada instante de síntesis según la fórmula 5.

$$\hat{h}(t) = \sum_{k=-L}^{L} \hat{A}_k\left(t_s^i\right) e^{j2\pi k f_o'(t_s^i)(t-t_s^i)} \quad (5)$$

Para obtener la frecuencia fundamental requerida, f_o', será necesario recalcular las amplitudes complejas de los armónicos sin que se modifique la envolvente espectral del sonido. Este ajuste de las amplitudes se lleva a cabo mediante una interpolación en función de las

amplitudes de los dos armónicos originales entre los que se sitúa el nuevo armónico. Los instantes de síntesis se calcularán en función de la frecuencia de *pitch* y de la duración que se deseen para el habla sintetizada. Una vez reconstruidas las partes armónicas de cada trama, se construirá la señal armónica mediante un proceso de solapamiento y suma.

La señal de ruido correspondiente se generará directamente por superposición y suma de las tramas de la parte de ruido que se han obtenido en el proceso de análisis. Se utilizarán las mismas marcas de síntesis que para la parte armónica, por lo que la duración de las dos señales será la misma.

La señal de voz final será la suma de las dos componentes, armónica y de ruido, generadas previamente.

3. Conversión texto-habla (CTH)

La conversión texto-habla se puede dividir en un proceso *off-line* y otro *on-line*. El proceso *off-line* parte de un corpus de voz en catalán que consta de 895 difonemas y 313 trifonemas que comprenden todas las posibles combinaciones fonéticas. Hay que destacar que cada difonema contiene dos fonemas enteros y cada uno de ello tiene 5 marcas de segmentación obtenidas durante el proceso de análisis. Estas marcas identifican el inicio y final del fonema, el inicio y final de la parte estable y su centro y se utilizan como referencia en el proceso de síntesis ya que el punto de concatenación será variable en función del contexto fonético de las unidades a sintetizar.

El proceso de análisis de dicho corpus se completa con el etiquetaje de las unidades, el marcaje automático del *pitch* y por último el cálculo de los parámetros de la parte armónica y de ruido del modelo híbrido utilizado.

El proceso *on-line* es el que genera el habla sintética, en el cual las unidades necesarias se concatenan y se modifican según la prosodia deseada. Gracias al esquema síncrono con el *pitch* que se utiliza tanto en el proceso de análisis como en el de síntesis, la modificación prosódica será sencilla. En primer lugar se deben generar los instantes de síntesis, que coincidirán con las marcas de *pitch* de síntesis para los tramos sonoros y unos valores interpolados para los segmentos no sonoros.

Después se tienen que generar las tramas de la parte armónica según la frecuencia fundamental deseada (fórmula 5), lo que supone la modificación de las amplitudes complejas de los armónicos correspondientes siguiendo el proceso descrito en el apartado 2.3.

El último paso, previo a la superposición y suma siguiendo los instantes de síntesis, es el ajuste de la duración de los segmentos según la prosodia deseada. Para llevar a cabo este proceso se deben eliminar o añadir tramas hasta aproximarse al máximo a la duración requerida. Este ajuste en la duración se consigue mediante una combinación lineal de las tramas situadas alrededor del punto de concatenación (límites de los difonemas), con la ventaja añadida de suavizar la transición en dicha unión.

4. Resultados

La primera valoración de resultados ha consistido en la comparación del sistema híbrido presentado con el anterior TD-PSOLA mediante la comparación de un mismo texto sintetizado con ambos métodos. Dicho texto está formado por nueve frases extraídas de la voz en *off* de un documental de televisión, de las que se ha obtenido la correspondiente información prosódica, que se utilizará como dato de entrada al sistema. El resultado de la comparación muestra una ligera mejoría del nuevo método respecto al sistema basado en TD-PSOLA. Hay que destacar que la síntesis del texto utilizado no requiere de grandes variaciones prosódicas y por lo tanto la técnica TD-PSOLA tiene un funcionamiento aceptable.

En un segundo estudio del funcionamiento del sistema, se parte de un texto formado por once frases con la información prosódica correspondiente a cuatro emociones (alegría, tristeza, rabia y miedo). Se han escogido estas cuatro emociones como resultado del estudio presentado en (Iriondo et al., 2000). Los primeros resultados comparativos, que se han llevado a cabo de manera informal, muestran una notable mejora para el caso de la síntesis del habla emocionada, en el cual las variaciones prosódicas (sobre todo de *pitch*) son importantes.

El próximo paso será integrar el sintetizador en un sistema de síntesis audiovisual, con el cual se podrá llevar a cabo una prueba de percepción y obtener resultados cuantitativos que permitan valorar la importancia de la síntesis combinada de voz y animación facial para la comprensión del mensaje y la simulación de expresión emocional. En la figura

2 se muestra el aspecto del locutor virtual en el que se va a integrar el nuevo sintetizador de voz.

Figura 2: Aspecto del locutor virtual

5. Conclusiones

En este artículos se ha presentado la mejora que supone la substitución de un sistema de síntesis basado en TD-PSOLA por otro basado en un modelo híbrido que añade una método de parametrización "armónico+ruido" de la señal de voz en el ámbito de la conversión texto habla concatenativa por difonemas. La principal ventaja que presenta el método propuesto es la capacidad de generar altas variaciones de *pitch* respecto la señal original consiguiendo un sonido natural. Esta característica es imprescindible en los sistemas de síntesis de habla emocionada.

Los siguientes pasos para la mejora del sistema de síntesis consisten en el perfeccionamiento de ciertos aspectos, como son la detección de la máxima frecuencia sonora y la parametrización de la parte de ruido. Además, en un futuro próximo se pretende integrar este sistema de síntesis en un conversor texto habla basado en selección de unidades con el cual se consiga mejorar la calidad segmental del habla generada.

Agradecimientos Este trabajo se ha realizado en parte con el apoyo del Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya mediante la beca 2000FI-00679 del DOGC 07/02/01.

Bibliografía

Alías, F. y I. Iriondo. 2001. Asignación automática de marcas de pitch basada en programación dinámica. *Procesamiento del Lenguaje Natural*, (27):225–231, September.

Dutoit, T. y H. Leich. 1996. MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13:435–440, June.

Iriondo, I., R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, D. Tena, y L. Longhi. 2000. Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. *Proceedings of the ISCA Workshop on Speech and Emotion*, páginas 161–166, September.

Moulines, E. y F. Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, December.

Rodríguez, A., P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, y L. Longhi. 1999. Modelización acústica de la expresión emocional en el español. *Procesamiento del Lenguaje Natural*, (25):159–166, September.

Stylianou, Y. 1996a. A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech. *IEEE Nordic Signal Processing Symposium*, September.

Stylianou, Y. 1996b. *Harmonic plus Noise Models for Speech combined with Statistical Methods for Speech and Speaker Modification*. Ph.D. tesis, Ecole Nationale Superieure des Telecommunications, Paris.

Stylianou, Y. 2001. Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis. *IEEE Transactions on Speech and audio Processing*, 9(1):21–29, January.

Stylianou, Y., T. Dutoit, y J. Schroeter. 1997. Diphone Concatenation using a Harmonic plus Noise Model of Speech. *Proceedings EUROSPEECH'97*, September.

Syrdal, A., Y. Stylianou, L. Garrison, A. Conkie, y J. Schroeter. 1998. TD-PSOLA versus Harmonic plus Noise Model in diphone based speech synthesis. *ICASSP-98*.