Hacia una síntesis concatenativa de alta calidad para aplicaciones de conversión texto-habla

Ignasi Iriondo, Josep Martí, Jaume Oliver, Roger Guaus, Helena Moure
{ iriondo | marti | joliver | rguaus | helena }@salleURL.edu
Dept. de Comunicacions i Teoria del Senyal. Enginyeria La Salle. Universitat Ramon Llull
Pg. Bonanova 8, 08022 Barcelona

RESUMEN

Este artículo describe nuevas líneas de investigación referentes a la síntesis concatenativa para la conversión texto habla. La técnica TD-PSOLA supuso un salto importante en cuanto a la mejora de la calidad de los sistemas anteriores. Ha sido un método válido para muchas aplicaciones pero es insuficiente para las nuevas necesidades en el campo de las telecomunicaciones, donde se requiere una síntesis más natural y agradable. Se hace referencia a nuevos enfoques de los sistemas de conversión texto habla, especialmente en dos aspectos, el método y la estrategia de síntesis. Respecto al método de síntesis se explican mejoras conseguidas con modificaciones efectuadas sobre el TD-PSOLA. En referencia a la estrategia de síntesis se describen las líneas de investigación actuales haciendo una especial mención de la necesidad de disponer de herramientas automáticas de segmentación y etiquetado de bases de datos de voz.

1. Introducción

El objetivo de la síntesis de habla consiste en la generación de información oral por parte de una máquina en un contexto de comunicación hombre-máquina. Sin embargo, los resultados obtenidos hasta el momento no permiten disponer de sistemas que generen un habla completamente natural. Para mejorar la calidad de los sistemas de conversión texto-habla (CTH) en cuanto a naturalidad, se tiene que avanzar en tres áreas: 1) análisis lingüístico, 2) modelado de la prosodia, y 3) modelos de síntesis de habla. Para conseguir síntesis de habla de alta calidad, las tres áreas mencionadas tienen la misma importancia.

En este artículo se describen líneas de investigación referentes a la síntesis concatenativa para aplicaciones CTH que han surgido recientemente. La técnica TD-PSOLA (Time Domain Pitch Syncronous OverLap and Add) [1] supuso un salto importante en cuanto a la mejora de la calidad de los sistemas anteriores, tales como síntesis basada en coeficientes de predicción lineal o en formantes. Durante esta década ha sido el sistema base para muchas aplicaciones CTH, pero el nivel de calidad no es suficiente para las aplicaciones que requieren una mayor naturalidad de la voz sintetizada. Por lo tanto, se dedica un primer capitulo para analizar las principales limitaciones de la técnica TD-PSOLA.

A continuación se describen posibles mejoras del TD-PSOLA. En esta parte se hace referencia, en primer lugar, al método MBR-PSOLA, indicando la filosofía seguida para mejorar algunas de las limitaciones del TD-PSOLA. En un segundo apartado, describimos un método propio para conseguir suavizar las transiciones entre uniones de forma que la evolución temporal del espectro no presente discontinuidades.

Otro punto fundamental en el diseño de un sistema de síntesis de alta calidad es la estrategia de síntesis que se utilizará, es decir, la naturaleza que tendrán las unidades de la base de datos (fonemas, difonemas, etc.) y el criterio de selección de dichas unidades (en función de información prosódica, fonética, de continuidad espectral, etc.). En este apartado se describe la síntesis por selección de unidades que se basa en la utilización de grandes bases de datos de voz con múltiples repeticiones de la misma unidad. Se hace referencia también a la necesidad de desarrollar herramientas de segmentación automática para el tratamiento de dichas bases de datos.

2. Limitaciones del TD-PSOLA

Este apartado pretende centrar los objetivos de una síntesis de alta calidad y sobre todo justificar las limitaciones del TD-PSOLA.

La aparición de la técnica TD-PSOLA [1], permitió dar un salto cualitativo en el campo de los sistemas CTH. Básicamente, las limitaciones de esta técnica de síntesis son consecuencia de que no utiliza un modelo de producción de voz, sino que trabaja directamente con la forma de onda temporal de las unidades grabadas. Mediante un análisis síncrono con el pitch, se utilizan ventanas 2T (siendo T el período fundamental), de forma que la síntesis de habla se realiza mediante la superposición y suma de las unidades de la base de datos. Las variaciones de pitch se realizan separando o juntando las tramas 2T. Las variaciones de duración se consiguen repitiendo o eliminando tramas (ver figura 1).

Los principales inconvenientes que presenta el algoritmo TD-PSOLA se pueden clasificar en tres grupos:

a) Variaciones prosódicas:
   - La modificación de pitch comporta una modificación en la duración del segmento, que se tiene que compensar.
   - La variación de duración sólo se puede realizar de forma cuantificada (resolución de un período de pitch).
   - El alargamiento de sonidos sordos mediante repetición de tramas, da lugar a ciertos clics metálicos.

b) Base de datos: Como consecuencia de no utilizar un modelo de producción de habla, las unidades de la base de datos no están parametrizadas, sino que se almacenan las muestras de la forma de onda. El tamaño de la base de datos es muy grande en comparación con otros sistemas de síntesis.

c) Concatenación: Se pueden producir discontinuidades a diferentes niveles entre las tramas final e inicial de las unidades a concatenar:
   - Discontinuidad en la fase
   - Discontinuidad en el pitch
   - Discontinuidad espectral

A pesar de estas limitaciones, la síntesis TD-PSOLA tiene la gran ventaja de proporcionar un habla sintética de una cierta calidad con un coste computacional muy bajo.

3. Mejoras respecto TD-PSOLA

Debido a las limitaciones detalladas en el apartado anterior, surge la necesidad de mejorar el método de síntesis. En este apartado se hace referencia a tres métodos de síntesis que intentan mejorar la síntesis TD-PSOLA.

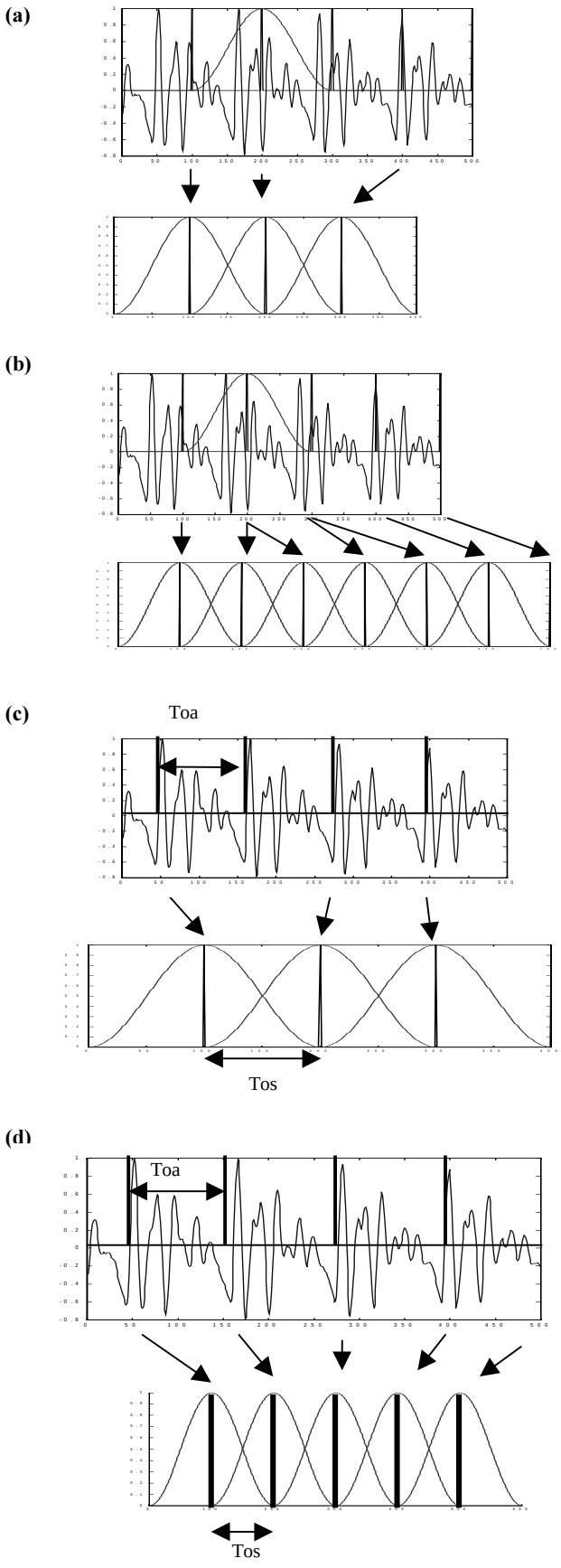

Fig. 1. Síntesis TD-PSOLA. (a) Disminución duración. (b) Aumento duración. (c) Disminución pitch. (d) Aumento pitch.

Se trata del algoritmo MBR-PSOLA y de un método desarrollado por nuestro grupo que suaviza la evolución temporal del espectro.

3.1 Algoritmo MBR-PSOLA

El algoritmo MBR-PSOLA, Multi-Band Resynthesis Pitch Synchronous OverLap and Add [2], tiene como objetivo principal solucionar los problemas que presenta el algoritmo TD-PSOLA en cuanto a concatenación de unidades. Los problemas de discontinuidad de fase, salto en el pitch y la discontinuidad de la envolvente espectral se reducen con un tratamiento a dos niveles. En primer lugar se trabaja con la base de datos que previamente se ha almacenado siguiendo el algoritmo TD-PSOLA. El tratamiento que se realiza consiste en analizar y resintetizar todos los segmentos sonoros de la base de datos siguiendo el modelo Multi-Band Excited (MBE) [3]. El hecho de volver a sintetizar los segmentos sonoros permite fijar la misma frecuencia fundamental y hacer una corrección en la fase. En segundo lugar, se realiza una interpolación lineal de tramas en el dominio temporal durante la fase de síntesis.

La resíntesis MBE sólo se tiene que efectuar una vez y únicamente sobre los segmentos sonoros, por lo tanto no se incrementa el coste computacional de la síntesis CTH.

Las mejoras introducidas con esta variante consiguen suavizar un poco las transiciones pero no son suficientes debido a que hay demasiada dependencia con el TD-PSOLA.

3.2 Mejoras en la evolución espectral

En el marco de la síntesis basada en difonemas, la concatenación de unidades se realiza por la parte estable de dos fonemas iguales. El problema principal radica en que las dos unidades a concatenar se han grabado en contextos diferentes. Por lo tanto, estas dos unidades estarán coarticuladas por los fonemas vecinos en el momento de la grabación y al concatenarlas, se producirá una discontinuidad en la evolución temporal del espectro del fonema. Esta discontinuidad se produce en cada unión y produce una degradación en el habla generada.

Este fenómeno tiene un mayor efecto en las uniones CV - VC debido a la mayor duración y estabilidad de las vocales. Se propone un sistema capaz de decidir el punto de unión entre dos difonemas [4].

La decisión del punto de unión vendrá dada por el tipo de coarticulación de las dos unidades a concatenar.

Se establecen dos posibilidades, la primera consiste en utilizar la totalidad de una de las dos vocales, uniendo por uno de los extremos. En el caso de no poder unir por uno de los extremos, se busca el punto de unión óptimo y se realiza una interpolación entre las tramas de las dos unidades.

Para poder unir por un extremo, las condiciones articulatorias tienen que ser muy favorables de forma que una de las dos consonantes CV - VC sea muy parecida a la consonante contextual del otro difonema. La cuantificación de esta medida de similitud se realiza mediante una función distancia que tendrá un valor pequeño si los espectros son semejantes y un valor alto si son diferentes. Si la distancia por la izquierda es pequeña, se utilizará la vocal derecha entera. En cambio, si la distancia por la derecha es pequeña, se utilizará toda la vocal izquierda.

Cuando la decisión no está clara, la unión se realiza por un punto intermedio mediante interpolación de tramas. La elección del punto de unión es variable y dependerá de los valores obtenidos anteriormente de las funciones distancia por la izquierda y por la derecha.

Con este método se consigue mejorar sensiblemente la continuidad en el habla sintética, aunque sigue teniendo las limitaciones del TD-PSOLA referentes a la variación prosódica y al tamaño de la base de datos.

4. Estrategias de síntesis.

Los sistemas tradicionales de síntesis CTH utilizan un conjunto preestablecido de unidades que permita la generación de cualquier mensaje oral. Las unidades más utilizadas han sido los difonemas y los trifonemas dado que permiten la concatenación por la parte más estable de los fonemas. Con este método se han conseguido altos niveles de inteligibilidad y una calidad aceptable para un determinado número de aplicaciones. En los últimos años se han desarrollado nuevas estrategias en el diseño y naturaleza de las bases de unidades de voz [5].

4.1 Selección de unidades

Estas nuevas estrategias parten de una gran base de unidades, segmentada y parametrizada de forma adecuada. En esta gran base de unidades de voz, habrá repeticiones de éstas,

pero que tendrán contextos fonéticos diferentes. Para generar un texto concreto se seleccionan las unidades que consiguen una mayor naturalidad en la voz sintetizada. La selección de las unidades se lleva a cabo mediante una función de coste que tiene en cuenta parámetros prosódicos e información del contexto fonético. Se escogen las unidades que mejor se adaptan a los requisitos de duración y pitch solicitados y que además consiguen disminuir la discontinuidad espectral con las unidades vecinas. Si para la elección de éstas solamente se tiene en cuenta información que garantice una alta continuidad espectral, la duración y el pitch se tienen que modificar posteriormente.

El hecho de utilizar grandes bases de datos comporta nuevas necesidades de carácter práctico como son:

- Disponer de herramientas de segmentación automática.
- Intentar reducir el tamaño en memoria de la base de datos mediante la utilización de unidades parametrizadas
- Diseñar algoritmos eficientes para seleccionar las unidades óptimas para generar un mensaje concreto.

4.2 Segmentación de bases de datos

La segmentación automática se hace imprescindible cuando se necesita generar una base de datos de voz de gran tamaño. Realizar esta tarea de forma manual resulta muy costoso y tiene el inconveniente de la diferencia de criterio si se realiza por personas diferentes. El objetivo de la segmentación automática consiste en generar un conjunto de marcas que delimiten todas las unidades de la base de datos a partir de los ficheros de voz y la transcripción fonética correspondiente. Las técnicas más utilizadas son las que se utilizan en el reconocimiento automático del habla.

El método de segmentación automática implementado utiliza los modelos ocultos de Markov (HMM) para modelar la voz. El proceso de segmentación automática tiene dos fases: 1) Fase de entrenamiento o generación de los HMM para cada fonema. 2) Segmentación automática utilizando los modelos de los fonemas.

La fase de entrenamiento parte de un pequeño conjunto de frases segmentadas de forma manual. Por cada fonema se genera un modelo de Markov izquierda-derecha [6].

Para la fase de segmentación se genera un modelo concatenando todos los modelos de los fonemas de la frase. Mediante el algoritmo de alineamiento de Viterbi se busca la secuencia de estados óptima. Recorriendo el camino de estados se establecen las marcas de segmentación con una precisión en muestras equivalente al tamaño de la trama utilizada.

5. Conclusiones

En este artículo se han descrito las actuales líneas de investigación que buscan conseguir una síntesis de alta calidad. En cuanto al método de síntesis, la técnica TD-PSOLA y sus variantes han alcanzado un tope de calidad que necesita de nuevos enfoques para ser superado. Últimamente ha aparecido una familia de métodos de síntesis que puede suponer un nuevo salto cualitativo en las aplicaciones CTH. Se trata de los modelos híbridos armónico/aleatorio.

Los modelos híbridos descomponen la señal de voz en una componente armónica más una componente ruidosa o aleatoria. Esta descomposición permite modificaciones en la señal consiguiendo una mayor naturalidad en el habla sintética.

Las dos componentes se separan en el dominio frecuencial mediante un parámetro que varía en el tiempo, la frecuencia sonora máxima [7]. Por debajo de esta frecuencia máxima se asume que se representará la componente armónica y por encima la componente ruidosa.

La componente armónica se modela como suma de sinusoides con amplitud y fase variables y frecuencia múltiple de la fundamental.

La componente ruidosa se describe frecuencialmente mediante un modelo autorregresivo (AR) variante en el tiempo.

Esta técnica permite mejorar diferentes aspectos del TD-PSOLA a cambio de aumentar el coste computacional.

Respecto a la estrategia de síntesis, la tendencia actual tiende a sustituir los sistemas basados en un conjunto mínimo de unidades de voz por grandes bases de datos con repetición de unidades que tienen diferentes características prosódicas y fonéticas. Por motivos de sencillez en el diseño, los primeros sistemas de este tipo utilizan como unidad base los fonemas.

6. Referencias

1. E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication,* vol. 9, pp. 453-467, Dec 1990.
2. T. Dutoit and H. Leich, “MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database”. *Speech Communication,* vol. 13, pp. 435-440, June 1996.
3. D. Griffin and J. Lim, ”Multiband-excitation vocoder”, *IEEE Trans. Acoust. Speech Signal Process*., vol. ASSP-64, pp 236-243. Feb 1988.
4. R. Guaus, J. Oliver, H. Moure, I. Iriondo, J. Martí. *Sintesis de Voz por concatenación de unidades: mejoras en la calidad segmental*. Acústica'98. Lisboa, Portugal. Set 1998
5. A. Hunt and A. Black, “Unit Selection in a concatenative speech synthesis system using a large speech database”, *IEEE Int. Conf. Acoust. Speech, Signal Process*., pp. 373-376. 1996.
6. L.R Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. of the IEEE, Vol.77, nº 2, Feb 1989.*
7. Y. Stylianou, “Concatenative Speech Synthesis using a Harmonic plus Noise Model”. *Proc. ESCA/COCOSDA Workshop on synthesis.* Jenolen Caves, Nov 1998.