Síntesis del habla

Helenca Duxans Barrobés
Marta Ruiz Costa-jussà

PID_00188071

UOC
Universitat Oberta de Catalunya

Índice

Introducción..... 5

Objetivos..... 6

1. Introducción a la síntesis del habla..... 7

2. Aplicaciones de los convertidores de texto a voz..... 9

3. Los convertidores de texto a voz..... 10
3.1. El módulo de procesamiento de lenguaje natural 10
3.1.1. Analizador morfosintáctico 11
3.1.2. Transcriptor fonético 11
3.1.3. Generador prosódico 12
3.2. El módulo de procesamiento de la señal 13

4. Síntesis por concatenación..... 14
4.1. Elementos de los sistemas de síntesis por concatenación 14
4.2. Unidades acústicas y construcción del corpus 15
4.3. Selección de las unidades 15
4.4. Concatenación 16
4.4.1. Concatenación en el dominio temporal 16
4.4.2. Concatenación en el dominio paramétrico 17

5. Modificaciones prosódicas..... 19
5.1. Técnicas basadas en TD-PSOLA 19
5.1.1. Modificaciones de altura tonal 19
5.1.2. Modificaciones de duración 21
5.1.3. Alternativas al algoritmo de TD-PSOLA 23

6. Medidas de calidad de la voz sintetizada..... 25

Introducción

En este módulo introducimos el concepto de síntesis del habla y las aplicaciones que tiene. En el estado actual de la tecnología, existe un abanico muy amplio de estrategias y algoritmos que se utilizan para crear voz sintética a partir de un texto escrito. La elección de la técnica que se debe utilizar depende de muchos factores: de la calidad de la voz que se requiere, del número de voces diferentes que se necesitan, de la cantidad de memoria que puede utilizar la aplicación, del tiempo de reacción del sistema, etc.

En este módulo nos centraremos en la estructura de los sistemas de conversión de texto a voz más utilizada en los sistemas comerciales actuales y en las técnicas de síntesis por concatenación en el dominio temporal.

¿Por qué se debe estudiar la síntesis del habla?

La interacción por medio del habla con los dispositivos que nos rodean requiere no solo que estos dispositivos nos entiendan (reconocimiento del habla), sino que nos puedan transmitir información. El medio de transmisión de esta información puede ser variado; por ejemplo, visual (puntos luminosos, texto, gráficas, imágenes o vídeos), acústico (sonidos o habla), háptico (vibraciones, temperatura, etc.) o una combinación de dos medios o más.

La síntesis del habla reúne todas las técnicas necesarias para transformar en voz cualquier información.

Objetivos

Al acabar de trabajar este módulo, deberéis ser capaces de lo siguiente:

1. Explicar el funcionamiento de un convertidor de texto a voz a alto nivel.

2. Identificar los módulos principales de un convertidor de texto a voz.

3. Relacionar los conceptos de coarticulación y de discontinuidades en los sistemas de síntesis por concatenación.

4. Calcular los nuevos instantes y ventanas de concatenación del algoritmo TD-PSOLA para realizar modificaciones prosódicas.

1. Introducción a la síntesis del habla

En un sentido muy amplio, la síntesis del habla se puede definir como todo proceso de creación de habla de manera artificial (por medios mecánicos, electrónicos, etc.).

Nota

La reproducción directa de grabaciones de audio no se considera síntesis del habla, puesto que no ha asociado ningún proceso de creación de contenido en el instante de la reproducción.

En este módulo nos centraremos en los sistemas digitales de síntesis actuales.

Bibliografía complementaria

Para una visión de la evolución histórica de los sistemas de síntesis del habla podéis consultar las referencias siguientes:

D. H. Klatt (1987). "Review of Text-to-Speech Conversion for English". *Journal Acoustical Society of America* (núm. 82, vol. 3, pág. 737-793).

P. Calliope (1989). *La parole et son traitement automatique* (pág. 410-414). París: Masson/CNET-ENST ("Technique et Scientifique des Télécommunications").

R. Lingaard (1985). *Electronic Synthesis of Speech* (pág. 1-17). Cambridge: Cambridge University Press.

Los sistemas de síntesis actuales se dividen en dos grandes grupos:

- **Sistemas de respuesta vocal o de texto restringido**
  - Finalidad: generación acústica de frases con una estructura predeterminada, en la que solo varía parte del vocabulario.
  - Ejemplo: avisos de los trayectos de los trenes ("Próxima estación: Palacio Real") o información telefónica ("El número solicitado es 902 141 141").
  - Técnica: funcionamiento basado en la concatenación de trozos de frase o palabras que se han grabado previamente de manera aislada.
  - Ventajas: la calidad de la voz generada es elevada, a pesar de que a veces se notan los puntos de unión de las diferentes grabaciones.
  - Inconvenientes: no sirven para aplicaciones en las que el vocabulario es muy variable (carteleras de cine, por ejemplo), puesto que incluir palabras nuevas en el vocabulario implica realizar nuevas grabaciones.
- **Convertidores de texto a voz**
  - Finalidad. Lectura de cualquier texto, con independencia del origen que tenga.
  - Ejemplo. Lectores de pantalla para personas con deficiencias visuales, entre otros.

  - Técnica. Funcionamiento basado en dos pasos: un análisis del texto dado y la generación del audio correspondiente.

  - Ventajas. Flexibilidad, puesto que tienen vocabulario ilimitado, y altamente configurables (por ejemplo, sexo y edad del locutor, y velocidad a la hora de hablar).

- Inconvenientes. En comparación con los sistemas de respuesta vocal, la complejidad técnica es muy superior.

El resto del módulo lo focalizaremos en describir los convertidores de texto a voz y las técnicas de procesamiento de audio que están asociadas a estos.

2. Aplicaciones de los convertidores de texto a voz

Una gran parte de las aplicaciones de los convertidores de texto a voz o *text-to-speech conversion system* (TTS) está relacionada con sistemas de diálogo, en los que trabajan junto con un ASR. Dentro de este tipo de aplicaciones destacamos la **atención telefónica** (servicios de información, de notificación de averías, de concertar cita, etc.) y las aplicaciones de **mando y control** (sistemas domóticos, interacción vocal con coches, PC, etc.), que ya introdujimos en el módulo "Reconocimiento automático del habla".

Los TTS también se utilizan mucho en aplicaciones de **ayuda a personas con necesidades especiales**. Los lectores de pantalla, junto con teclados adaptados, han permitido el acceso al mundo de la informática e Internet a las personas con dificultades visuales. Para las personas con dificultades en la producción del habla que les afectan a la inteligibilidad, existen dispositivos con teclados predictivos o de construcción rápida de frases, que les permiten sintetizar todo aquello que quieren expresar.

Otro campo de aplicación de los TTS es el **aprendizaje de idiomas**. En este caso, los estudiantes de idiomas extranjeros pueden sintetizar cualquier frase o palabra para aprender una pronunciación y entonación correctas. Los sistemas de **traducción automática voz-voz**, a pesar de encontrarse todavía en fase precomercial, constituyen un nuevo campo de aplicación.

3. Los convertidores de texto a voz

Los convertidores de texto a voz tienen como entrada un texto y como salida el audio correspondiente.

Aunque existen diferentes técnicas para realizar esta conversión, todos los TTS comparten la misma arquitectura que se muestra en la figura 1:

Arquitectura TTS

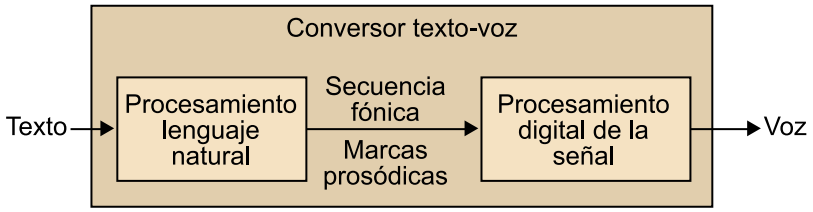


Figura 1. Arquitectura de los convertidores de texto a voz

Como se aprecia en esta figura, la arquitectura de un TTS está formada por dos módulos: uno de procesamiento de lenguaje natural y otro de procesamiento digital de la señal.

El **módulo de procesamiento de lenguaje natural** es el encargado de transformar el texto de entrada (texto plano, HTML, etc.) en una transcripción fonética con marcas prosódicas.

Marcas prosódicas

Información sobre cuándo se producen los cambios audibles en el tono, la forma de los contornos melódicos, el volumen y el ritmo del habla. Por ejemplo, no tiene la misma prosodia una oración enunciativa que una interrogativa.

Podéis encontrar una definición más completa de prosodia en el módulo "Introducción al habla".

El **módulo de procesamiento digital de la señal** genera el audio descrito en la entrada; por ello también se denomina *módulo de síntesis acústica*.

En los próximos apartados veremos con más detalle cada uno de los dos bloques de un TTS.

3.1. El módulo de procesamiento de lenguaje natural

El módulo de procesamiento de lenguaje natural o *natural language processing* (PLN) es el encargado del análisis del texto de entrada y de tratarlo para que lo pueda leer un sistema de síntesis.

Un sistema completo de procesamiento de lenguaje natural para un TTS consta, mayoritariamente, de tres etapas:

- Un analizador morfosintáctico
- Un transcriptor fonético

- Un generador prosódico

3.1.1. Analizador morfosintáctico

El analizador morfosintáctico procesa el texto para dejarlo limpio (eliminación de etiquetas, viñetas, etc.) y extrae la estructura del texto, es decir, la organización de las frases según las relaciones de dependencia que se establecen entre las palabras. También aporta información sobre la categoría gramatical de las palabras o *part of speech* (POS) y normaliza el texto expandiendo abreviaciones, acrónimos o transcribiendo los números, entre otras operaciones.

Categoría gramatical

Por *categoría gramatical* se entiende la función que tiene la palabra dentro de la frase. Por ejemplo, nombre/sustantivo, pronombre, adjetivo, verbo o adverbio.

La figura siguiente muestra un ejemplo de análisis morfosintáctico para la frase *La CE defiende el criterio de la ONU*:

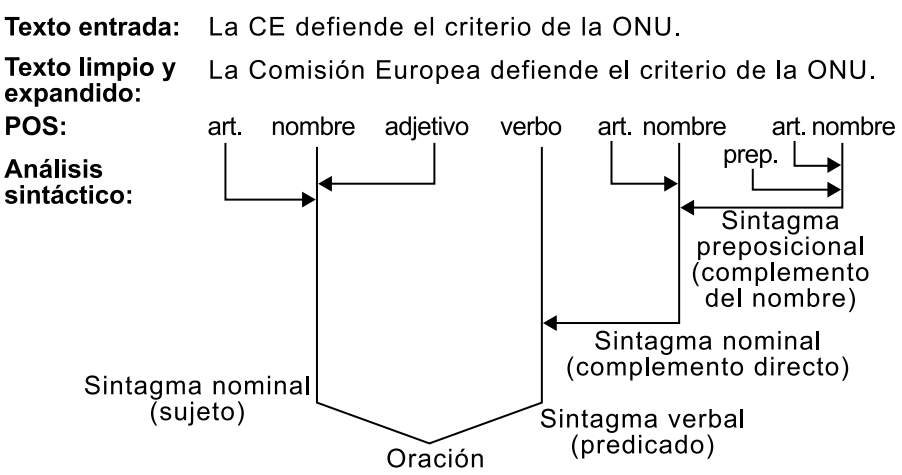


Figura 2. Ejemplo de análisis morfosintáctico de la frase *La CE defiende el criterio de la ONU*

Hay tres razones principales por las cuales los TTS incorporan un analizador morfosintáctico:

1. Para saber cómo se ha de expandir el texto no normalizado (*CE* y *ONU* en el ejemplo anterior; otros casos son los números o las siglas, como *PSOE*, *PP*, *IU* y *CC. OO.*).

2. Porque la transcripción fonética puede depender de la categoría gramatical de la palabra.

3. Porque la prosodia depende de la sintaxis.

3.1.2. Transcriptor fonético

El transcriptor fonético transforma el texto limpio y expandido en una secuencia de fonemas, teniendo en cuenta la información morfosintáctica asociada a cada palabra y las relaciones fonéticas que se establecen entre unas y otras para el idioma elegido. Se debe pensar, por ejemplo, que hacer una conversión tex-

to-fonemas en castellano es muy trivial en comparación con el inglés, puesto que en castellano la relación es prácticamente 1:1 (es decir, una letra tiene en la mayoría de los casos solo un fonema asociado). Por ejemplo, la vocal *a* en castellano tiene solo un fonema asociado ([a]), mientras que en inglés puede tener varios ([a], [ɛ], etc.).

Podéis ver la transcripción en el alfabeto fonético internacional de las frases siguientes:

Catalán: *és una bona infermera* [e'Zunəbo'niɱfərme'rə]

Castellano: *es una buena enfermera* [esu'naβwe'naeMferme'ra]

Inglés: *she is a good nurse* [ʃi' i'z ə gʊ'd nə'rs]

3.1.3. Generador prosódico

El generador prosódico es el encargado de generar de manera automática una descripción de los contornos melódicos y los patrones de ritmo de las locuciones. La prosodia tiene una función muy clara en la comunicación verbal, que es situar el foco de atención, transformar frases declarativas en interrogativas o exclamativas y transmitir el estado emocional del locutor.

Información prosódica

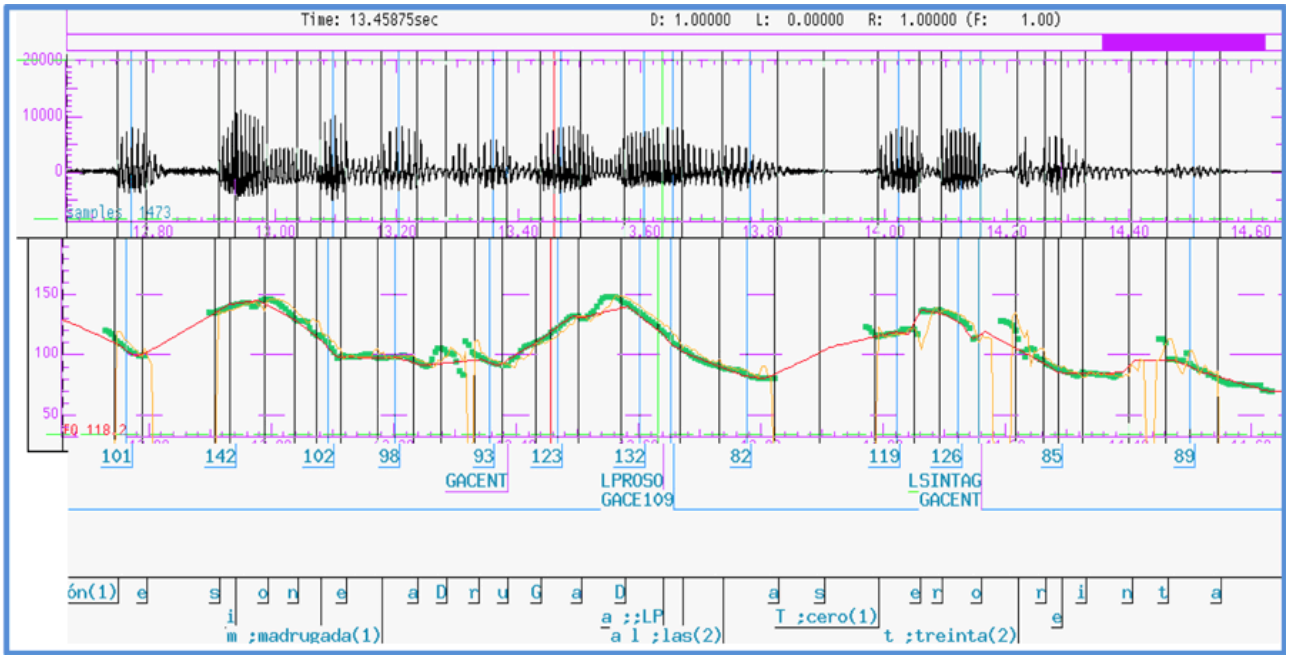


Figura 3. Ejemplo de un segmento de voz (gráfico superior) con la curva melódica (gráfico del medio) y duración de cada fonema (gráfico inferior)

Bibliografía complementaria

Los métodos utilizados en el módulo de procesamiento de lenguaje natural son técnicas de lingüística computacional e inteligencia artificial. Si queréis más información, podéis consultar las referencias siguientes:

M. A. Martí; J. Llisterri (2002). *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita*. Barcelona: Edicions de la Universitat de Barcelona.

T. Dutoit (1997). *An Introduction to Text-to-Speech Synthesis* (pág. 37-176). Kluwer Academic Publishers.

3.2. El módulo de procesamiento de la señal

El módulo de procesamiento de la señal es el encargado de transformar la información simbólica de fonemas y marcas prosódicas, proporcionada por el módulo de procesamiento de lenguaje natural, en habla. Por lo tanto, es el módulo que genera la señal acústica.

Las primeras estrategias de síntesis acústica que surgieron imitaban el sistema de producción humano (síntesis articulatoria) o las características acústicas de la señal (síntesis por formantes). En la actualidad, la mayoría de los TTS comerciales utilizan la síntesis por concatenación en el módulo de procesamiento de la señal, por su simplicidad y por la calidad de la voz que genera.

En el apartado siguiente explicaremos con más detalle la síntesis por concatenación.

Bibliografía complementaria

Para ampliar la información sobre los sistemas de síntesis articulatoria y síntesis por formantes, podéis consultar la referencia siguiente:

X. Huang; A. Acero; A. Hon (2001). *Spoken Language Processing* (pág. 796-803). Englewood Cliffs, NJ: Prentice Hall PTR.

Los TTS transforman cualquier texto en voz.

Los TTS están formados por dos módulos: el módulo de procesamiento de lenguaje natural y el módulo de procesamiento digital de la señal.

El módulo de procesamiento de lenguaje natural procesa el texto de entrada para extraer la transcripción fonética y genera las características prosódicas más adecuadas para este texto.

El módulo de procesamiento digital de la señal convierte la secuencia de fonemas y la información prosódica, proporcionadas por el módulo de lenguaje natural, en una onda acústica.

La técnica más utilizada de generación de la onda acústica es la síntesis por concatenación, a pesar de que existen otras técnicas, como por ejemplo la síntesis por formantes y la síntesis articulatoria.

4. Síntesis por concatenación

La síntesis por concatenación consiste en poner, uno detrás de otro, trozos cortos de grabaciones de un mismo locutor para reproducir la transcripción fonética con las características prosódicas requeridas.

Reflexión

Si las locuciones se sintetizan a partir de voz real, ¿la calidad de los sistemas por concatenación es similar a la voz natural? Escuchemos el audio de ejemplo: *tts.mp3*. Claramente, la calidad de la voz sintética no es igual a la voz real.

La calidad de la voz sintetizada no es igual a la de la voz real. El principal motivo de la pérdida de calidad de los sistemas de concatenación es la coarticulación. El fenómeno de coarticulación provoca que, cuando se concatenan dos segmentos de voz que no eran adyacentes en la grabación original, se produzcan **discontinuidades** en los puntos de concatenación, y por lo tanto se dé una pérdida de calidad/naturalidad.

Coarticulación

Adaptación del comienzo y final de un sonido a los sonidos adyacentes para que la voz suene fluida. Para más información al respecto, podéis consultar el módulo "Introducción al habla".

Hay dos tipos de discontinuidades:

- **Las discontinuidades espectrales**. Son saltos bruscos en la distribución frecuencial de la señal, debidos al hecho de que los formantes de las dos unidades no coinciden en el punto de concatenación.

- **Las discontinuidades prosódicas**. Se generan principalmente cuando el tono (f_0) o volumen de cada unidad no coincide en el punto de concatenación.

Los sistemas de síntesis por concatenación intentan minimizar estos dos tipos de discontinuidades.

4.1. Elementos de los sistemas de síntesis por concatenación

Los sistemas de síntesis por concatenación están compuestos por una base de datos (denominada *corpus de unidades acústicas*) y tres bloques funcionales: un bloque de selección de unidades acústicas, un bloque de modificación prosódica y un bloque de concatenación.

Sistemas de síntesis por concatenación

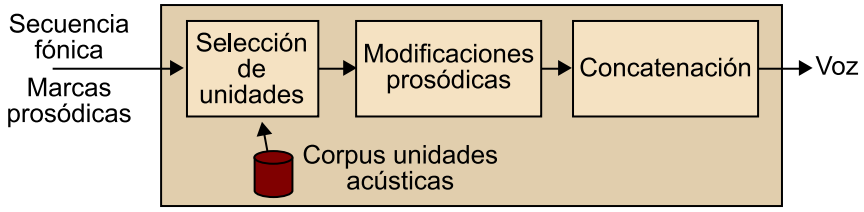


Figura 4. Módulo de procesamiento digital de la señal (o módulo de síntesis acústica) de un TTS basado en síntesis por concatenación

La generación de la voz empieza cuando el bloque de selección elige los segmentos de voz que se concatenarán, de entre todos los disponibles en el corpus de unidades acústicas. Esta selección se realiza teniendo en cuenta la transcrip-

ción fonética y el patrón melódico requeridos. A continuación, se modifica la prosodia de cada segmento si es muy diferente de la prosodia que se quiere. En el último paso, se concatenan los segmentos.

El bloque de modificaciones prosódicas no se encuentra siempre en todos los TTS basados en concatenación. Si la dimensión del corpus de unidades acústicas es muy grande, los segmentos seleccionados tienen una prosodia muy parecida a la requerida, y por lo tanto no es necesario modificarla.

Ved también

Las técnicas de modificación prosódica las introduciremos en el apartado "Modificaciones prosódicas".

4.2. Unidades acústicas y construcción del corpus

Uno de los requisitos principales en la elección de las unidades acústicas que se han de utilizar en un TTS es que se generen pocas discontinuidades en la concatenación.

Una de las unidades más utilizadas en los TTS son los **difonemas**, puesto que tienen el comienzo y final en zonas estables acústicamente y, por lo tanto, proporcionan una distorsión en la concatenación más baja que otras unidades.

Difonema

Recordad que ya hemos visto la definición de difonema en el módulo "Introducción al habla".

Un difonema es la unidad acústica que empieza en medio de la zona estable de un fonema y acaba en medio de la zona estable del fonema siguiente.

Para construir el corpus se debe diseñar un conjunto de frases que incluyan todas las unidades acústicas que se querrán sintetizar y hacer grabaciones con un locutor profesional. Una vez grabado el material, se etiquetan todas las grabaciones junto con la información prosódica que está asociada a estas (la altura tonal, la duración, la energía, etc.). Finalmente, si la unidad acústica que utilizaremos es el difonema, se deben indicar los límites de cada segmento de señal perteneciente a un difonema para guardarlo en el corpus junto con la información prosódica (altura tonal, duración, etc.).

El proceso de creación del corpus de un TTS requiere mucha intervención manual: la elección del locutor, las grabaciones y la revisión de los etiquetados. El proceso de etiquetado es una tarea semiautomática; esto significa que hay una primera fase de etiquetado automático y una segunda fase de revisión manual.

Voz de un TTS

El locutor que graba el corpus de unidades lingüísticas *cede* la voz al TTS. Cabe señalar que no todo el mundo puede ser locutor de un corpus, puesto que no todos los locutores generan voces sintéticas de buena calidad. Por este motivo, se contrata a locutores profesionales.

4.3. Selección de las unidades

El objetivo de este bloque es elegir la sucesión de unidades acústicas óptima para sintetizar la descripción fonética proporcionada y que, además, se ajuste a la prosodia requerida.

La elección de las unidades para una secuencia de fonemas no es trivial, puesto que en la base de datos normalmente existe más de una instancia para cada unidad, cada una de las cuales posee una prosodia diferente. La técnica más utilizada para realizar esta selección es una búsqueda utilizando el algoritmo de Viterbi, que minimiza una función de coste que representa la distorsión introducida.

La **función de coste** está formada por dos términos:

- El **coste de unidad** representa la diferencia entre los valores de prosodia y de contexto que tiene la unidad elegida respecto a los valores requeridos. Normalmente se calcula de manera proporcional a la diferencia entre las características prosódicas (f_0 o $\log(f0)$, duración, energía, etc.) de la unidad elegida y la descripción requerida de entrada al módulo.

- El **coste de concatenación** representa la pérdida de calidad debida a la unión de dos unidades. Se calcula como la distancia entre los parámetros espectrales de dos unidades adyacentes.

La búsqueda prioriza la selección de unidades adyacentes en las grabaciones originales, siempre que las diferencias prosódicas no sean muy notables. Por lo tanto, se reducen los puntos de concatenación y se evitan pérdidas de calidad.

4.4. Concatenación

La concatenación es el último paso para sintetizar una locución. Su objetivo es unir la secuencia de segmentos acústicos seleccionados previamente de modo que disminuyan las discontinuidades que se dan en los puntos de concatenación.

Para reducir las discontinuidades se utilizan técnicas de suavización. Según el dominio en el que se apliquen, se distingue entre concatenación en el dominio temporal y concatenación en el dominio paramétrico.

4.4.1. Concatenación en el dominio temporal

La concatenación en el dominio temporal se basa en la idea de superponer y sumar o, en inglés, *overlap-and-add* (OLA). Esta técnica se aplica en dos fases: el análisis (cuando se definen los segmentos de voz que se superpondrán) y la síntesis (cuando se realiza la suma).

En la fase de análisis se construyen ventanas de voz multiplicando los segmentos acústicos seleccionados por ventanas superpuestas. Los instantes en los que se hacen estas multiplicaciones se denominan *instantes de análisis* t_a (podéis ver la parte superior de la figura 5).

Una práctica habitual, denominada *time domain pitch synchronous overlap-and-add* (TD-PSOLA), es situar los puntos de análisis separados un período de altura tonal y elegir ventanas de análisis de duración dos períodos de altura tonal.

En la fase de síntesis, los segmentos de voz se sitúan en los instantes de síntesis t_s que les corresponden (dependiendo de qué altura tonal tiene la voz que queremos crear) y se suman (podéis ver la parte inferior de la figura 5).

La figura 5 muestra un ejemplo de TD-PSOLA para la concatenación de dos unidades acústicas:

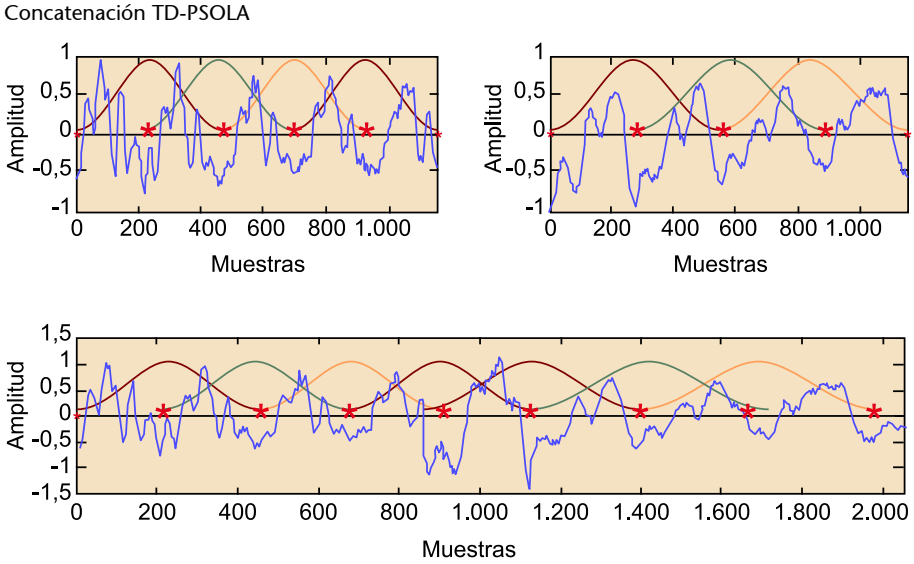


Figura 5. Concatenación TD-PSOLA de dos unidades acústicas. Arriba: los dos segmentos temporales que se han de concatenar. Abajo: el resultado de la concatenación. Los puntos de análisis y síntesis se han indicado con un asterisco rojo.

En los intervalos que se encuentran alrededor de los puntos de concatenación, el valor de la forma de onda resultante es la interpolación lineal de las dos unidades acústicas.

4.4.2. Concatenación en el dominio paramétrico

Para llevar a cabo la concatenación en el dominio paramétrico, primero se han de transformar los segmentos acústicos seleccionados previamente a la representación paramétrica según el modelo de voz utilizado. Los modelos de voz más utilizados para la concatenación son dos: el modelo fuente-filtro y el modelo armónico o la extensión armónico + ruido de este modelo.

Para que la transición entre los dos segmentos acústicos que se han de concatenar sea suave, en la frontera entre los dos segmentos se interpolan los parámetros del modelo de voz de cada segmento. Finalmente, a partir de los nuevos parámetros interpolados se genera la voz.

La síntesis por concatenación consiste en crear la voz sintética mediante la unión de segmentos cortos pregrabados de un mismo locutor.

Todos los esfuerzos de los sistemas de concatenación se centran en minimizar las discontinuidades espectrales y prosódicas en los puntos de concatenación.

Los principales componentes de los sistemas de concatenación son los siguientes:

- Un corpus de unidades acústicas. Las unidades más utilizadas son las que tienen el comienzo y el final en zonas estables frecuencialmente (por ejemplo, los difonemas).
- Un bloque de selección de las unidades que minimiza el coste de unidad (diferencia entre la unidad seleccionada y la especificada) y el coste de concatenación (medida de las discontinuidades introducidas por la unidad en los puntos de concatenación).
- Un modificador de la prosodia (bloque opcional, dependiendo de la dimensión del corpus).
- Un bloque de concatenación en el dominio temporal o paramétrico.

Las técnicas de concatenación más habituales en el dominio temporal se basan en el algoritmo OLA.

Bibliografía complementaria

Si queréis profundizar en los modelos de voz y la utilización de estos modelos en la síntesis por concatenación, podéis consultar las referencias siguientes:

T. Dutoit (1997). *An Introduction to Text-to-Speech Synthesis* (pág. 229-248). Kluber Academic Publishers.

X. Huang; A. Acero; A. Hon (2001). *Spoken Language Processing* (pág. 290-333). Englewood Cliffs, NJ: Prentice Hall PTR.

5. Modificaciones prosódicas

Las técnicas de modificación prosódica permiten modificar la duración, la altura tonal y la energía de segmentos de voz, sin modificar el mensaje. Estas técnicas se utilizan en el bloque de modificación prosódica de un TTS, si lo tiene (recordemos que es opcional). También se pueden utilizar para cambiar la prosodia de cualquier grabación de voz original; por ejemplo, para ajustar la duración a un valor requerido o para ajustar las características de la voz a unos valores requeridos (por ejemplo, para subir o bajar el tono de voz o el volumen).

Para modificar la energía de la voz, es decir, el volumen de la voz, solo se ha de multiplicar la señal por un valor que se quiera (denominado también *ganancia*). En cambio, la modificación de la duración y la altura tonal requieren manipular la señal y la calidad de la voz resultante se degrada.

Existen, igual que para la concatenación, dos aproximaciones para modificar la altura tonal y la duración: técnicas basadas en la modificación de los parámetros de la señal de voz y técnicas basadas en OLA. Estas últimas técnicas son de las más utilizadas, sobre todo las basadas en TD-PSOLA, puesto que son sencillas de implementar y generan una voz modificada de calidad aceptable.

En el apartado siguiente describimos el funcionamiento de las técnicas de modificación prosódica basadas en TD-PSOLA.

5.1. Técnicas basadas en TD-PSOLA

El algoritmo de modificación prosódica en el dominio temporal TD-PSOLA o *time domain pitch synchronous overlap-and-add* consiste en modificar los instantes de síntesis y la secuencia de ventanas que se han de concatenar dependiendo del tipo de modificación que se requiera.

5.1.1. Modificaciones de altura tonal

En el algoritmo TD-PSOLA, las ventanas de análisis y síntesis se sitúan separadas un período de altura tonal. Por lo tanto, si se quiere modificar la altura tonal, solo se han de reubicar los instantes de síntesis según el nuevo período de altura tonal. En la figura 6 se muestra un ejemplo de modificación de altura tonal para un factor de 1,25 (multiplicación de 1/1,25 = 0,8 del período de altura tonal).

Modificación de altura tonal basada en TD-PSOLA

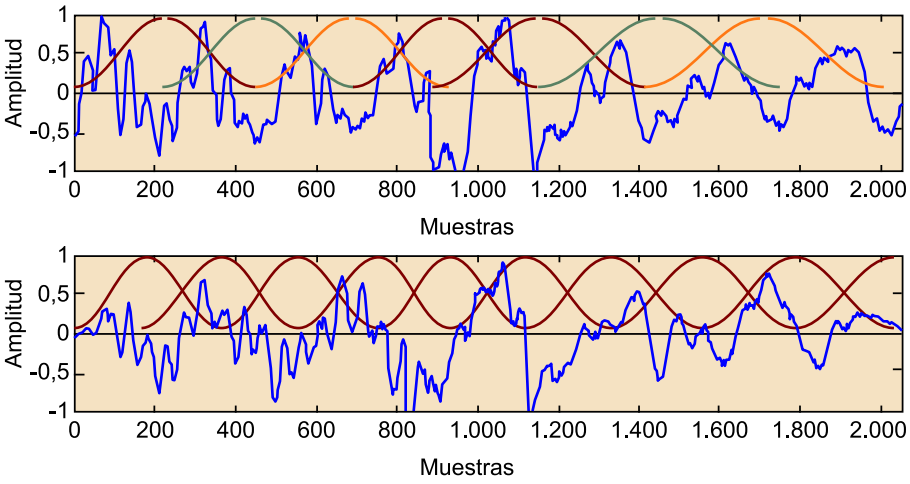


Figura 6. Modificación de TD-PSOLA de altura tonal por un factor constante de 1,25

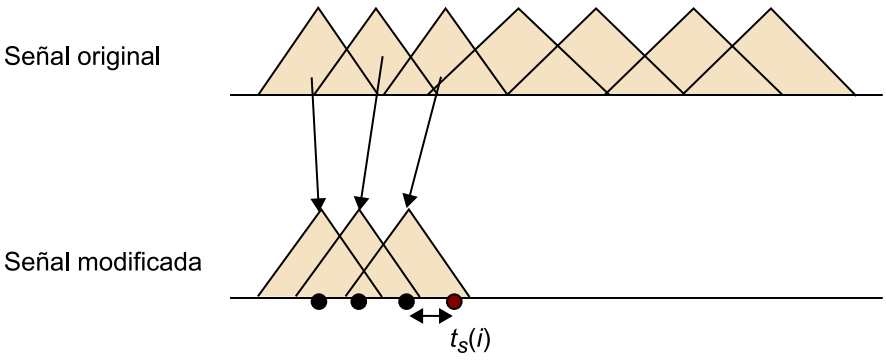
Podéis escuchar un ejemplo para valorar la calidad y el efecto de la modificación de altura tonal: "locució original" y "locució modificada".

Como observamos en la figura anterior, para que no se produzca modificación de la duración total del segmento de voz en la síntesis, se repiten o se eliminan ventanas de análisis.

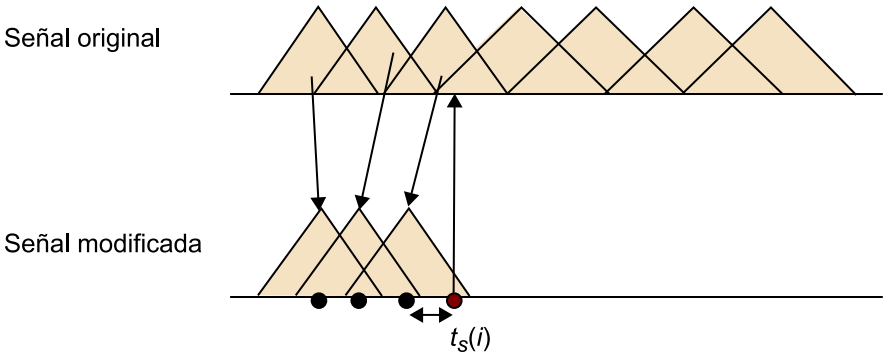
De una manera más esquemática, el algoritmo de TD-PSOLA para modificar la altura tonal consiste en los pasos que presentamos a continuación. Hemos ilustrado cada paso del algoritmo con un pequeño ejemplo gráfico en el que representamos de manera esquemática una señal original que se ha de modificar e indicamos las ventanas de análisis y la señal modificada que se va generando:

1. Determinar el nuevo instante de síntesis $t_s(i)$ según el factor de modificación $\alpha(i)$, del período de altura tonal $T_{pitch}(i)$ y del instante de síntesis $t_s(i-1)$:

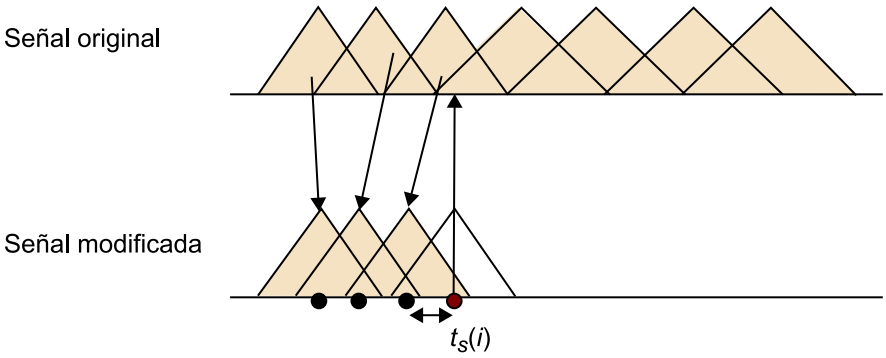
$$t_s(i) = t_s(i-1) + T_{pitch}(i)*\alpha(i)$$



2. Seleccionar la ventana de análisis más cercana al instante de síntesis:



3. Situar la ventana de análisis seleccionada en el instante de síntesis:



4. Una vez se han situado todas las ventanas en los instantes de síntesis correspondientes, se debe realizar la suma. En este ejemplo hemos reducido el período de altura tonal; por lo tanto, el sonido modificado resulta más agudo que el original.

Los cambios de altura tonal que permite el algoritmo de TD-PSOLA se encuentran entre valores de 0,5 (reducir a la mitad) y 2 (doblar), porque si no, la calidad queda muy degradada y se puede perder el mensaje del segmento de voz.

Para las partes sordas de la señal de voz, no se hace ninguna modificación.

5.1.2. Modificaciones de duración

Las modificaciones de duración se realizan repitiendo o eliminando ventanas de análisis, manteniendo siempre la altura tonal original.

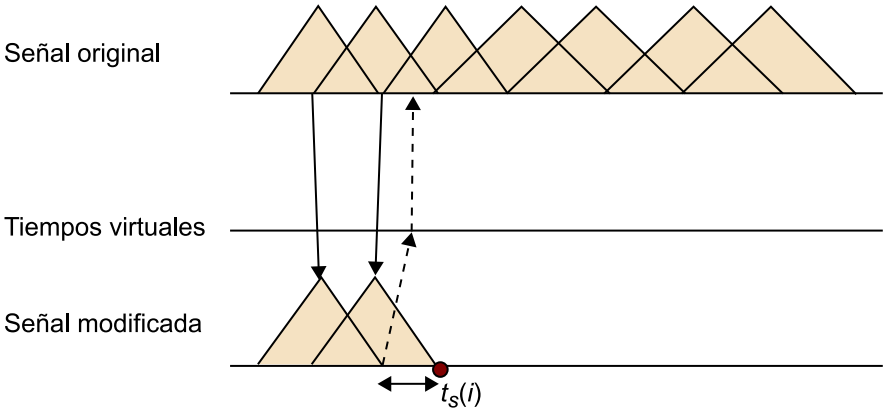
> Podéis escuchar la "locución original" y la "locución modificada" para valorar la calidad y el efecto de una modificación de duración por un factor constante de 1,25.

Para seleccionar las ventanas que se sitúan en cada punto de síntesis, se debe construir una secuencia de instantes virtuales. Los pasos del algoritmo son los que describimos a continuación. Como antes, hemos ilustrado cada paso

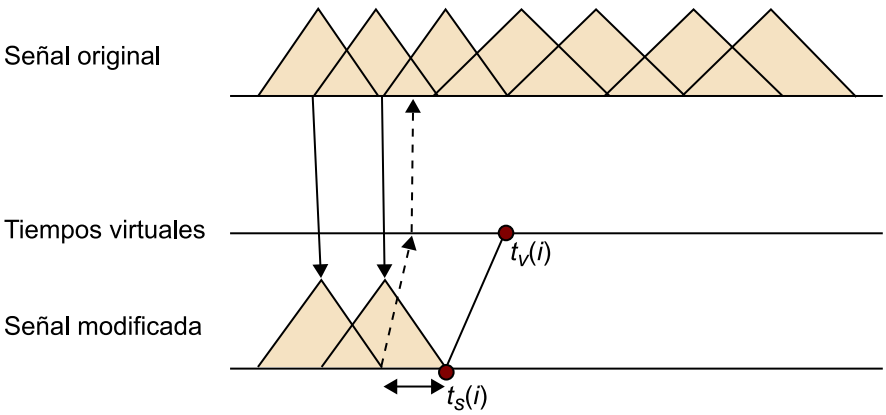
del algoritmo con un pequeño ejemplo gráfico en el que representamos de manera esquemática una señal original que se ha de modificar e indicamos las ventanas de análisis y la señal modificada que se va generando:

1. Determinar el nuevo instante de síntesis $t_s(i)$ según el período de altura tonal $T_{pitch}(i)$ y del instante de síntesis previo $t_s(i-1)$:

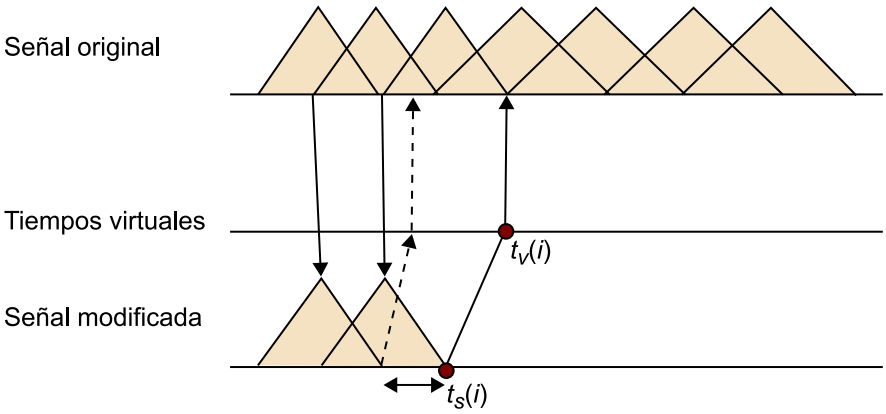
$$t_s(i) = t_s(i-1) + T_{pitch}(i)$$



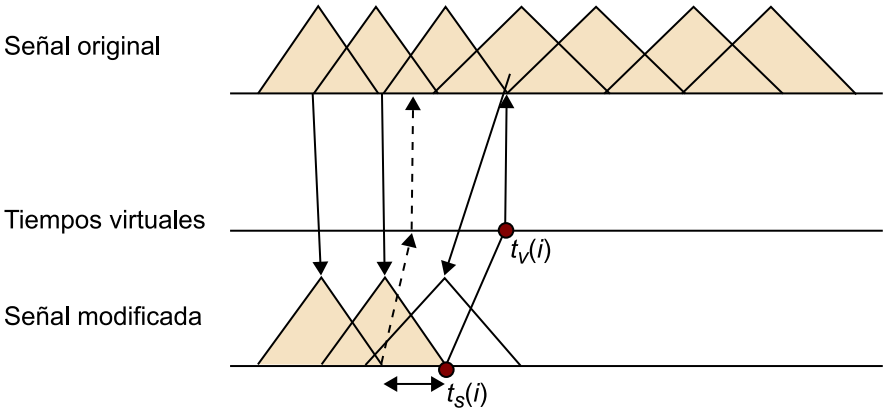
2. Determinar el nuevo instante virtual $t_v(i)$ según el instante de síntesis $t_s(i)$ y el factor de modificación duración $\beta(i)$: $t_v(i) = t_s(i-1) + (t_s(i) - t_s(i-1)) * \beta(i)$:



3. Seleccionar la ventana de análisis más cercana al instante virtual:



4. Situar la ventana de análisis seleccionada en el instante de síntesis:



5. Una vez se han situado todas las ventanas en los instantes de síntesis correspondientes, se debe realizar la suma. En este ejemplo hemos reducido la duración de la señal original.

Los cambios de duración que permite el algoritmo de TD-PSOLA se encuentran entre valores de 0,25 y 2, porque si no, la calidad queda muy degradada y se puede perder información relevante para el mensaje transportado por la voz.

Para las partes sordas de la señal de voz, los instantes de análisis y síntesis se sitúan en intervalos regulares (habitualmente de unos 10 milisegundos).

5.1.3. Alternativas al algoritmo de TD-PSOLA

La estimación de los instantes de análisis síncronos con la altura tonal es un procedimiento complicado y poco preciso, que provoca desajustes de fase, altura tonal y amplitud en los puntos de concatenación de la síntesis y, por lo tanto, efectos audibles en la señal modificada.

Por esta razón surgieron todo un conjunto de técnicas de la familia PSOLA que intentan minimizar las discontinuidades en los puntos de concatenación, como por ejemplo FD-PSOLA (*frequency domain pitch synchronous overlap-and-add*) y MBR-PSOLA (*multiband resynthesis pitch synchronous overlap-and-add*).

Bibliografía complementaria

Si queréis profundizar en otros algoritmos de modificación prosódica basados en PSOLA, podéis consultar las referencias siguientes:

T. Dutoit (1997). *An Introduction to Text-to-Speech Synthesis*. Kluber Academic Publishers.

Juego de herramientas de código abierto para construir un TTS:

Festival

6. Medidas de calidad de la voz sintetizada

Para saber cuál es la calidad de la voz generada por un TTS, la respuesta que parece más evidente es escuchar la voz sintetizada.

Ejemplo

Para tener una idea de la calidad actual de los TTS, podéis consultar los ejemplos comerciales siguientes:

Loquendo

Verbio

Nuance

Aun así, escuchando solo un ejemplo no se puede evaluar la calidad de un TTS, ni hacer una comparación entre diferentes sistemas, puesto que los TTS no se comportan igual con todas las frases de entrada.

Los tests más utilizados para evaluar los TTS son los siguientes:

- Test de inteligibilidad. Los evaluadores escuchan monosílabos y han de marcar qué han entendido de una lista de opciones muy parecidas acústicamente (por ejemplo: *mueve*, *nueve*). También se pide a los evaluadores que realicen transcripciones de frases sintetizadas sin sentido (por ejemplo, *la pelota come tres coches sube*).

- Test de calidad o *mean opinion score* (MOS). Se evalúa la calidad de cada locución del 1 al 5 (1 = *bad*; 2 = *poor*; 3 = *fair*; 4 = *good*; 5 = *excellent*).

Para comparar diferentes TTS entre sí, se debe asegurar que los evaluadores han pasado los mismos tests.

Los TTS comerciales actuales tienen una calidad muy buena para ser aceptados por el mercado para la mayoría de las aplicaciones que requieren una voz neutra. Hoy en día, los principales esfuerzos de investigación y comercialización están centrados en dos aspectos: la reducción de requisitos de memoria para poder introducir un TTS en cualquier dispositivo y la introducción de expresividades (por ejemplo, emociones o reír) en la voz sintetizada.

Los TTS transforman cualquier texto en voz.

Los TTS están formados por dos módulos: el módulo de procesamiento de lenguaje natural (encargado de generar la transcripción fonética y las marcas prosódicas) y el módulo de procesamiento digital de la señal (encargado de generar la onda acústica).

La técnica más utilizada de generación de la onda acústica es la síntesis por concatenación, que consiste en unir segmentos cortos pregrabados de un mismo locutor para formar locuciones. Todos los esfuerzos de los sistemas de concatenación están centrados en minimizar las discontinuidades espectrales y prosódicas en los puntos de concatenación.

Uno de los algoritmos más utilizados para realizar la concatenación y las modificaciones prosódicas de la voz es el TD-PSOLA. En la fase de análisis, se segmenta la voz original en ventanas encabalgadas de longitud dos veces el período de altura tonal y a la vez situadas en cada período de altura tonal. En la fase de síntesis se construye la señal sumando las ventanas encabalgadas que hay en cada instante de síntesis.

La calidad de la voz generada por un TTS es bastante buena para que haya múltiples aplicaciones comerciales.