(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 401 014**

(21) Número de solicitud: 201131569

(51) Int. CI.:

G01L 13/02 (2006.01)

(12) SOLICITUD DE PATENTE A2

(22) Fecha de presentación:

28.09.2011

(43) Fecha de publicación de la solicitud:

16.04.2013

(71) Solicitantes:

TELEFONICA, S.A. (100.0%)
Gran Vía, 28
28013 MADRID ES

(72) Inventor/es:

ARMENTA LOPEZ DE VICUÑA, Ana;
ESCALADA SARDINA, Jose Gregorio y
RODRIGUEZ CRESPO, Miguel Angel

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

(54) Título: **MÉTODO Y SISTEMA PARA LA SÍNTESIS DE SEGMENTOS DE VOZ**

(57) Resumen:

Método y sistema para la síntesis de segmentos de voz.

La presente invención propone un nuevo método y sistema de síntesis de voz. La invención introduce el uso de ventanas asimétricas en el tiempo de síntesis, facilitando por tanto una mejor adaptación a los cambios prosódicos, y reduciendo la distorsión e incorpora innovaciones en la estrategia de colocación de puntos de inicio y en la estrategia de modificaciones prosódicas.

ES 2 401 014 A2

DESCRIPCIÓN

Método y sistema para la síntesis de segmentos de voz

CAMPO TÉCNICO

La presente invención se refiere, en general, a tecnologías de voz. Más específicamente, se refiere a técnicas de procesamiento de señales de voz digital que se usan, entre otras aplicaciones, en convertidores de texto a voz.

DESCRIPCIÓN DE LA TÉCNICA ANTERIOR

Muchos sistemas de conversión de texto a voz actuales se basan en la concatenación de unidades acústicas tomadas de grandes bases de datos con multitud de unidades acústicas almacenadas, que se han grabado previamente. Los sistemas de este tipo se conocen como sistemas TTS (*Text to Speech*) basados en corpus. Este enfoque proporciona el nivel de calidad requerido para el uso de convertidores de texto a voz en muchas aplicaciones comerciales (principalmente en la generación de información hablada a partir de texto en sistemas interactivos telefónicos, pero también en una cantidad creciente de contenidos multimedia generados automáticamente para su difusión y en Internet).

Como se ha dicho, los sistemas de texto a voz (TTS) basados en corpus se basan en la selección de unidades de grandes bases de datos que contienen muchos ejemplos de diferentes combinaciones de sonidos que difieren en su contexto fonético, prosodia, posición en la palabra y oración. La elección óptima de estas unidades según un criterio de costo mínimo (costos por unidad y costos de concatenación) reduce la necesidad de realizar cambios a las unidades, y mejora enormemente la calidad y naturalidad de voz resultante. Pero no es posible eliminar totalmente la necesidad de modificar y concatenar las unidades de voz grabadas previamente, puesto que los corpus son finitos y no pueden garantizar una cobertura completa para sintetizar de manera natural cualquier oración. Por lo tanto siempre permanecerá la necesidad de concatenar sonidos procedentes de diferentes segmentos de voz.

Las causas posibles de discontinuidad y defectos en la voz sintética son de diversos tipos:

1. La diferencia en las características del espectro de la señal en los puntos de concatenación: frecuencias y anchos de banda de las unidades formantes, forma y amplitud del envolvente espectral.

2. Pérdida de coherencia de fase entre las tramas de voz que están concatenadas. Éstas también pueden verse como desplazamientos relativos inconsistentes de la posición de las tramas de voz (ventanas) en ambos lados de un punto de concatenación. La concatenación entre tramas incoherentes provoca una desintegración o dispersión de la forma de onda que se percibe como una pérdida significativa de calidad. La voz resultante es no natural: mezclada y confundida.

3. Diferencias prosódicas (entonación y duración) entre las unidades grabadas previamente y la prosodia objetivo (deseada) para la síntesis de una unidad de voz.

Aunque la concatenación de unidades acústicas evita el difícil problema de modelar la producción de voz completamente humana, surge otro problema básico: cómo controlar la prosodia de los segmentos seleccionados que van a concatenarse, y cómo realizarlo sin complicaciones.

Por este motivo, los convertidores de texto a voz habitualmente emplean diversos métodos de procesamiento de señal que permiten modificar la prosodia de los segmentos de voz que van a concatenarse, y sintetizar una voz natural continua. Pero esta modificación debe degradar lo menos posible la señal original. La modificación de señal fue indispensable en los primeros sistemas de texto a voz (TTS), con pequeños segmentos de voz (por ejemplo, difonos) y un número relativamente pequeño de unidades (normalmente una unidad por identidad de difono). En estos sistemas la necesidad de realizar modificaciones (y grandes modificaciones) a las unidades es muy alta.

Existen diferentes métodos de representación y modificación de la señal de voz que se han usado en TTS.

Los métodos basados en ventanas de solapamiento y adición de la señal de voz en el dominio del tiempo (métodos PSOLA, “Pitch Synchronous Overlap and Add”) disfrutan una amplia aceptación y difusión. El más clásico de estos métodos se describe en “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using dyphones” (E. Moulines y F. Charpentier, Speech Communication, vol. 9, págs. 453-467, diciembre de 1990). Este tipo de algoritmos se conocen como algoritmos síncronos de “pitch” o de detección de “pitch”, ya que las tramas de señal de voz (ventanas) se obtienen de una manera síncrona con el periodo fundamental (que en este ámbito técnico se asimila al concepto de pitch, similar al concepto de “tono” en español). La ventanas de análisis deben centrarse en el momento de cierre de la glotis (GCI, “Instantes de cierre glotal”; *Glottal Closure Instants*) u otros puntos identificados dentro de cada periodo de la señal, que deben etiquetarse cuidadosa y consistentemente, para evitar alteraciones de puntos de unión de fase. El marcado o ubicación de estos puntos es una tarea laboriosa que no puede ser completamente automática (requiere ajustes), y que afecta al funcionamiento del sistema. La

modificación de la duración y frecuencia fundamental (F0) se lleva a cabo por medio de la inserción o borrado de tramas, y el alargamiento o estrechamiento de la misma (cada trama de síntesis es un periodo de la señal, y el desplazamiento o distancia entre dos tramas sucesivas es la inversa de la frecuencia fundamental). Como en todos los algoritmos síncronos de ”pitch”, también se relacionan las transformaciones de F0 y de duración (una modificación de la F0 implica una modificación en la duración). El único mecanismo para modificar la duración de alófono independientemente de la F0 (en cierta medida) es la replicación o borrado de trama. Puesto que los métodos PSOLA no incluyen un modelo explícito de la señal de voz, es difícil el trabajo de interpolación de las características espectrales de la señal en el punto de concatenación.

El método MBROLA (solapamiento y adición de resíntesis de multibanda; *Multi-Band Resynthesis Overlap and Add*) descrito en “Text-to-Speech Synthesis based on a MBE re-synthesis of the segments database” (T. Dutoit y H. Leich, Speech Communication, vol. 13, págs. 435-440, 1993) trata el problema de la falta de coherencia de fase en las concatenaciones sintetizando una versión modificada de las partes sonoras de la base de datos de voz, forzándolas a tener una fase y F0 determinadas (idénticas en todos los casos). Pero este proceso afecta la naturalidad de voz.

También se han propuesto métodos LPC (*Linear Predictive Coding*; codificación predictiva lineal) para síntesis de voz. Estos métodos limitan la calidad de voz que supone un modelo de único polo. El resultado depende en gran parte de si la voz original de referencia se adapta mejor o peor que las presuposiciones del modelo. Estos métodos a menudo presentan un problema con voces femeninas e infantiles.

También se han propuesto modelos de tipo sinusoidal, en los que la señal de voz se representa mediante una suma de componentes sinusoidales. Los parámetros de modelo sinusoidal pueden realizar de forma bastante directa e independiente de ambos la interpolación de parámetros tales como modificaciones prosódicas. En términos de garantizar la consistencia de acierto de puntos de fase, se han elegido algunos modelos para manipular una estimación de los momentos de cierre de la glotis (un proceso que no siempre da buenos resultados). En otros casos se ha asumido simplificar suposiciones que consideran una fase mínima (lo que afecta la naturalidad de voz en algunos casos, provocándoles percibir más vacío y almacenado en memoria intermedia).

Modelos sinusoidales han incorporado diferentes enfoques para resolver el problema de coherencia de fase. Por ejemplo, analizar la voz con ventanas que se mueven según la F0 de la señal, pero no necesita enfocarse en GCI. Estas tramas se sincronizan posteriormente en un punto común basado en información desde el espectro de fase de la señal, sin afectar la calidad de voz. Se aplica la propiedad de la transformada de Fourier en que añadir un espectro de componente lineal es equivalente al desplazamiento de fase de la forma de onda en el dominio del tiempo. Se fuerza la primera señal armónica que es un valor de fase resultante 0, y el resultado es que la voz de todas las ventanas está coherentemente enfocada sobre la forma de onda, independientemente de en qué punto particular en un periodo de la señal originalmente enfocada. Por tanto, las tramas corregidas pueden combinarse de manera coherente en la síntesis.

En el 2009, la patente española P200931212 fue presentada por Telefónica para un método de análisis, modificación y síntesis de voz, con aplicación en el dominio TTS. Este método es un modelo sinusoidal modificado, capaz de mantener la coherencia de fase de la señal, y obtener una concatenación más suave y un mejor control prosódico. Éste garantiza calidad y suavidad cuando existe una modificación prosódica y/o concatenación de segmentos desde contextos diferentes. Pero este modelo es de muy alto consumo de CPU en análisis, y de consumo moderado de CPU en síntesis (como la fase, la f0 y la amplitud tienen que volver a calcularse para cada pico de cada trama). También mantiene información muy detallada acerca de cada segmento de voz grabado previamente, e incluso con optimización y compresión de estos datos, las existencias acústicas (archivos que almacenan los parámetros de las tramas de voz de las diferentes voces o los diferentes hablantes sintéticos del TTS) son muy grandes. Y para un rápido tiempo de respuesta, toda esta información debe mantenerse en la memoria principal.

Estos problemas no son impedimentos en una aplicación “desktop” de usuario único, o no tanto en una plataforma de intercambio telefónico que da servicio a un sistema de respuesta de voz interactiva. Pero esta técnica no es muy adecuada para un dispositivo de baja capacidad, similar a un teléfono móvil. Así que se ha investigado una aproximación más ligera con el fin de resolver estos problemas.

Las soluciones más clásicas se basan en el algoritmo PSOLA del TD (*Time Domain*; dominio del tiempo). Este es un modelo de síntesis muy rápido, y el tamaño de los archivos de datos es también pequeño. Pero ha resultado que este método tiene problemas con transformaciones prosódicas moderadamente grandes, tanto en la F0 como en la duración, produciendo una degradación de calidad. El control de duración replicando o borrando tramas, puesto que no existe modelamiento espectral explícito, también puede producir anomalías en la voz sintética.

Otro problema encontrado frecuentemente en la literatura con este tipo de modelos síncronos de pitch en el dominio del tiempo, es la necesidad de una localización de inicio muy precisa en la fase de análisis. La mejor calidad se obtiene con localización de inicio manual, pero esta solución no es factible para grandes bases de datos usadas en sistemas de TTS modernos.

Se han propuesto sistemas automáticos para esta tarea. Algunos necesitan una grabación auxiliar hecha con un electrolaringógrafo, que produce una señal directamente relacionada con el pulso glotal, y es más fácil detectar los instantes de sincronía o “epochs”. Otras soluciones trabajan directamente sobre la señal de voz, pero la calidad no es perfecta, y se necesita una revisión manual.

Por estos motivos, se ha desarrollado una nueva solución, con requisitos muy pequeños de CPU y memoria, y con una calidad tan buena como cualquiera de los sistemas anteriores.

SUMARIO DE LA INVENCIÓN

La invención propone un nuevo método y sistema de síntesis de voz. La invención introduce el uso de ventanas asimétricas calculadas y aplicadas en el momento de síntesis, facilitando por tanto una mejor adaptación a los cambios prosódicos, y reduciendo la distorsión, e incorpora innovaciones en la estrategia de colocación de puntos de inicio (conocidos en el campo técnico como “onset points” y la estrategia de modificaciones prosódicas.

Los principales objetivos de la invención son:

- Reducir la carga de CPU de la síntesis de voz en un sistema de TTS
- Reducir el tamaño de los archivos de datos que representan al hablante sintético
- Reducir la distorsión cuando trata las modificaciones prosódicas
- Recuperar exactamente la forma de onda de voz cuando no se llevan a cabo modificaciones prosódicas ni concatenación

Con el fin aliviar el problema de la dependencia de colocación exacta de inicio en este tipo de métodos síncronos de periodo fundamental, se ha desarrollado un nuevo módulo para calcular automáticamente la colocación de inicio.

En un primer aspecto, se propone un método para la síntesis de señal de voz, en el que cada alófono que va a reproducirse en la señal de voz sintetizada tiene un valor objetivo deseado de duración y un valor objetivo deseado de frecuencia fundamental, denominado F0 objetivo, y en el que la señal de voz que va a sintetizarse se aparta de las unidades de señal de voz grabadas previamente de un hablante de referencia, estando cada unidad de señal de voz compuesta por una secuencia de tramas de señal de voz, denominadas tramas originales, teniendo cada trama original una frecuencia fundamental F0, denominada F0 original, y en la que, dada la secuencia de alófonos que va a reproducirse, se selecciona una secuencia de tramas de señal de voz originales correspondiente a dicha secuencia de alófonos, comprendiendo el método las siguientes etapas:

a) Asignar una F0 objetivo a cada una de las tramas originales de la secuencia seleccionada de tramas originales, basada en la F0 objetivo del alófono correspondiente, siendo el periodo objetivo asignado a cada trama 1/F0 objetivo de la trama.

b) Generar la señal de voz, comprendiendo esta etapa:

b1) Modificar la secuencia de tramas originales, enventanando dicha secuencia de tramas originales estando las ventanas centradas en el punto de separación entre cada dos tramas consecutivas, siendo las ventanas asimétricas, calculándose la longitud de la ventana de manera independiente para ambas tramas consecutivas situada cada una a un lado del punto donde se centra la ventana, es decir, siendo la longitud del ala derecha de la ventana el periodo objetivo de la trama situada a la derecha del punto en el que se centra la ventana y siendo la longitud del ala izquierda de la ventana, el periodo objetivo de la trama situada a la izquierda del punto en el que se centra la ventana.

En otro aspecto, se presenta un sistema que comprende medios adaptados para llevar a cabo el método según cualquier reivindicación anterior.

Finalmente, se presenta un programa informático que comprende medios de código de programa informático adaptados para llevar a cabo el método descrito anteriormente.

Para un entendimiento más completo de la invención, sus objetos y ventajas, puede tenerse referencia a la siguiente memoria descriptiva y a los dibujos adjuntos.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Para completar la descripción y con el fin de proveer un mejor entendimiento de la invención, se proporciona un conjunto de dibujos. Dichos dibujos forman una parte integral de la descripción e ilustran una realización preferida de la invención, que no debe interpretarse como restrictiva del alcance de la invención, sino

más bien como un ejemplo de cómo puede realizarse la invención. Los dibujos comprenden las siguientes figuras:

La figura 1 representa un diagrama de bloques general de un sistema de texto a voz.

La figura 2 muestra un diagrama de bloques en el módulo de creación de base de datos de voz según una de las realizaciones de la invención.

La figura 3 muestra una realización ejemplar con las etapas principales del esquema de localización de puntos de inicio.

La figura 4 muestra un esquema del algoritmo de transformación prosódica según una de las realizaciones de la invención.

Las figuras 4a, 4b y 4c muestran respectivamente una vista detallada y ampliada de las partes a, b y c respectivamente de la figura 4.

La figura 5 muestra un diagrama de bloques con una representación esquemática del algoritmo de solapamiento y adición síncrono con periodo fundamental clásico.

La figura 6 muestra un esquema clásico para sintetizar la voz con el algoritmo de solapamiento y adición síncrono de pitch en el que se modifica F0 de 1/T0 (original) a 1/T (objetivo)

La figura 7 muestra un diagrama de bloques con una representación esquemática de la generación de tramas según una de las realizaciones de la presente invención.

Los números de referencia y símbolos correspondientes en las diferentes figuras se refieren a partes correspondientes a menos que se indique lo contrario.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

Con el fin de presentar la invención innovadora desarrollada, un esquema general de los sistemas de TTS puede ayudar a entender sus funciones. El sistema de TTS (10) se presenta en la figura 1. La mayoría de estos módulos van a llevar a cabo las mismas tareas y en la misma forma que se conoce comúnmente en la técnica anterior. Los tres módulos en los que se enfoca la presente invención, es decir, los que van a incluir algunos cambios innovadores son el módulo de creación de base de datos de voz, el módulo de transformación prosódica y el módulo de síntesis de señal de voz (en líneas punteadas en la figura 1)

Un sistema de TTS recibe como entrada una información de texto (texto sencillo, o quizás enriquecido con marcas en un idioma similar al SSML) 11, y su objetivo es producir una señal de voz sintética (19) tan natural e inteligible como sea posible, correspondiente a la lectura en voz alta de una conferencia por parte de un ser humano. Para lograr este objetivo, el TTS tiene diferentes módulos, que tratan los diferentes aspectos de la tarea de lectura. Habitualmente, cada uno de estos módulos tiene un archivo asociado con un idioma o información específica del hablante.

Una breve descripción de estos módulos:

- Procesamiento lingüístico (12): Recibe el texto de entrada, y lo procesa con el fin de extraer o generar tanta información como sea posible. La secuencia de tareas llevadas a cabo es:
    - detección de sentencia
    - tokenización (división en partes fundamentales como palabras, símbolos…)
    - normalización
    - expansión de números, abreviaciones, fechas..., traducirlas a una secuencia de palabras
    - etiquetado de parte de voz
    - silabación
    - asignación de acentuación
    - análisis sintáctico o análisis sintáctico superficial

Este módulo genera como salida una secuencia de letras correspondientes a la lectura del texto de entrada, enriquecido con toda la información disponible. La mayoría de esta información procede de reglas dependientes del lenguaje, recopiladas y aisladas en la base de datos de información lingüística (14).

- Generación prosódica (13): Este módulo recibe la secuencia de letras del módulo previo y debe generar la

secuencia de sonidos (alófonos) que van a producirse, cada alófono con un valor de duración y una F0 (frecuencia fundamental) asociados al contorno (en algunos sistemas, hay también un contorno de energía prosódico). Para esto, se usa información prosódica general (15). La secuencia de tareas llevadas a cabo es:

- Inserción de pausa automática: Decide los puntos en los que deben realizarse pausas. Pueden estar marcados de manera ortográfica y de manera no ortográfica, reforzando siempre el significado de la oración. Este módulo sigue estadísticas y reglas tanto dependientes del hablante como dependientes del lenguaje.

- Conversión de grafema a fonema: Decide qué sonido (alófono o fonema) corresponde a cada grafema (letra) en la oración de entrada. Sigue habitualmente reglas dependientes del lenguaje.

- Asignación de la duración: Para cada alófono (también silencios), decide su duración, basándose en su tipo, la información de acentuación, su contexto, su posición en la oración, etc. Habitualmente se basa en estadísticas dependientes del hablante.

- Asignación de F0: Basándose en toda la información disponible (acentuación, sílabas, etiquetas de parte del discurso, tipo de oración, como enunciativa, interrogativa, exclamativa...), decide un contorno de F0 para la oración. Habitualmente, este contorno asigna tres valores (inicial, medio y final) de F0 a los alófonos relevantes (núcleo de sílaba, primer y último alófono de la oración, etc.). Si a un alófono no se le ha asignado F0, sus valores se obtendrán mediante la interpolación lineal a partir de los otros alófonos. Estos contornos pueden ser dependientes de reglas del hablante o estadísticos o dependientes de lenguaje.

Así se obtiene en la salida de este módulo, una secuencia de alófonos (sonidos), cada uno de ellos con un valor objetivo de duración, y un contorno objetivo de F0 (por ejemplo, tres valores de F0, al comienzo, en el centro y al final del alófono). En la descripción de las transformaciones prosódicas, estos valores se denominarán 'objetivo', ya que son los valores que el módulo de generación prosódica espera que se produzcan en la señal de voz sintética, en oposición a los valores 'originales' en los segmentos de voz grabados (obtenidos mediante el módulo de creación de base de datos de voz), y probablemente diferentes de los valores conseguidos en realidad finalmente en la señal de voz generada. La secuencia de alófonos puede tener también alguna información adicional, procedente del procesamiento lingüístico, como acentuación, o límites de sílabas.

- Módulo de síntesis (113): Este módulo recibe la secuencia de alófonos, con los valores objetivo de F0 y duración (y cualquier otra información adicional), y genera la mejor salida de voz sintética. La mayoría de los TTS modernos están accionados por unidad, lo que significa que la voz sintética se genera a partir de grabaciones de voz (organizadas como unidades, habitualmente difonos, la última mitad de un alófono y la primera mitad del siguiente alófono) de un hablante tomado como referencia. La mayoría de estos sistemas están también basados en corpus, habiendo múltiples ejemplos de cada unidad, grabados con diferentes contextos y prosodias.

    - Selección de unidad (16): A partir de la secuencia de alófonos, y el inventario de unidades de voz grabadas (base de datos de voz 110), encuentra la secuencia óptima de unidades para sintetizar la voz deseada (o, dicho de otro modo, encuentra la secuencia óptima de tramas para sintetizar la voz deseada, porque, tal como se explicará posteriormente, cada unidad está formada por una secuencia de tramas). En el caso de sistemas basados en corpus, este submódulo usa funciones de coste para ponderar los diferentes parámetros (acentuación objetivo frente a acentuación original, F0 objetivo frente a F0 original, continuidad de las unidades, etc.) para seleccionar la mejor secuencia de candidatos. En la salida de este módulo, se representa la voz como una secuencia de alófonos con los valores prosódicos objetivo, y una secuencia de unidades que hará realidad la voz, con la información original o grabada.

En este punto puede ser útil una breve introducción a la creación de base de datos de voz (111). De manera externa (y previa) al TTS, una vez seleccionado el hablante de referencia, se graba un corpus de voz amplio con su voz (112). A este corpus se le anota la información la secuencia de alófonos, los tiempos de comienzo y fin de todos ellos, el contorno de F0 (denominado F0 original) para todas las grabaciones (generadas habitualmente de manera automática con cualquier algoritmo de detección de periodo fundamental o pitch, y revisadas y corregidas manualmente), y cualquier otra información disponible (información de acentuación, secuencia de palabras, etiquetas de parte del discurso, flexiones, etc.). Toda esta información, con la realización acústica de cada alófono, se organiza en unidades. Cada unidad consiste en un fragmento breve de voz (habitualmente la última mitad de un alófono, y la siguiente mitad del siguiente). Según el modelo de síntesis del módulo de síntesis, esta información acústica de cada unidad se codifica como una secuencia de tramas. En el modelo descrito en esta invención, las tramas corresponden a periodos de señal de voz, de modo que la duración de la trama es igual a la longitud del periodo, la inversa de la F0 local (habitualmente, los algoritmos de detección de pitch proporcionan una versión suavizada de esta F0 local). El número de tramas en una unidad es igual al número de periodos en esa unidad.

- Transformación prosódica (17): Según las características del modelo de síntesis, pueden imponerse

los valores prosódicos objetivo sobre las unidades seleccionadas más o menos directamente. Obviamente, cualquier transformación con respecto a los valores prosódicos originales provocará una degradación de la calidad y naturalidad de la voz sintética, y la sensibilidad de esta degradación a los diferentes parámetros dependerá del modelo de síntesis implementado, de modo que habitualmente este módulo dependerá del módulo de síntesis de señal de voz. En la salida de este submódulo, la secuencia de unidades tendrá los valores prosódicos objetivo finales, según los cuales se generará la voz sintética.

- Síntesis de señal de voz (18): Este módulo toma la secuencia de unidades con los valores prosódicos objetivo finales, y genera la voz sintética. Con el fin de poder modificar (si fuera necesario) los valores prosódicos originales de las unidades, y concatenar suavemente cuando dos unidades consecutivas proceden de dos archivos de grabación diferentes, este submódulo se basa en un modelo de síntesis usado comúnmente, como los descritos en la técnica anterior. Estos modelos suelen estar directamente relacionados con los esquemas de codificación de voz: el módulo de creación de base de datos de voz corresponde a la fase de codificación, en la que la voz se representa como una secuencia de parámetros del módulo, y la síntesis de señal de voz corresponde a la fase de decodificación, en la que la secuencia de parámetros (modificada si fuera necesario según las transformaciones prosódicas) se transforma de nuevo en voz. Los modelos usados deben tener una representación explícita de parámetros prosódicos, de modo que puedan modificarse fácilmente.

Los módulos afectados por la presente invención son (en líneas discontinuas en la figura 1)

- **Creación de base de datos de voz**: Tal como se describió previamente, este módulo toma las grabaciones de voz originales, las organiza según la información lingüística (identidad de fonemas, acentuación, información de sílabas, POS, análisis sintáctico...), y codifica la señal de voz con un modelo de voz que permite modificaciones prosódicas y concatenación. Esta información se estructura en unidades. Cada unidad normalmente representa la última mitad de un alófono, y la primera mitad del siguiente. En cada unidad, la información acústica se representa como los parámetros de codificación de varias tramas. Puesto que el modelo de síntesis mejorado en esta innovación es un modelo de solapamiento y adición de dominio del tiempo, cada trama consiste en las muestras de voz de un periodo de voz, más la F0 (también denominada F0 local para distinguirla de la F0 suavizada), más la F0 suavizada predicha por el algoritmo de detección de periodo fundamental. La frecuencia fundamental F0 puede tener diferentes definiciones dependiendo del modelo usado. La F0 local se define como la inversa del periodo de trama. El periodo de trama vendrá dado preferiblemente por la separación entre los puntos de inicio (longitud de trama), es decir, la F0 local cambia para cada trama y es la inversa de la separación entre los puntos de inicios (la ubicación de estos puntos es una de las tareas principales de este módulo y se explicará más adelante). La F0 suavizada es una estimación de la F0 obtenida mediante los algoritmos de detección de “pitch” clásicos, que se aplica habitualmente a segmentos de voz más largos. La F0 suavizada se procesa habitualmente para eliminar las fluctuaciones locales (por eso se denomina suavizada). Así que puede decirse que la F0 local es la F0 más precisa (cuando se usa el término frecuencia fundamental F0, habitualmente se refiere a la F0 local), depende de las ubicaciones de los inicios y tiene una base de trama (es decir, se calcula trama por trama). La F0 suavizada es una estimación de la F0 local que se obtiene a través de técnicas de detección de periodo fundamental y que se aplica a un grupo de tramas.

- **Transformación prosódica**: Este módulo debe modificar los parámetros prosódicos originales de las tramas de voz (concretamente F0 y duración, aunque en los modelos síncronos de pitch ambos parámetros están relacionados) y ajustarlos a los valores deseados (objetivo). Estos valores objetivos se han especificado a nivel de alófono por el módulo de generación prosódica, así que este módulo debe distribuirlos a nivel de trama.

- **Síntesis de señal de voz**: Este módulo toma las tramas de voz que se han organizado en la base de datos de voz, codificadas con el modelo de voz usado en la creación de base de datos de voz. El módulo las modifica según la prosodia final decidida en el módulo de transformación prosódica, y las concatena con el fin de construir una nueva señal de voz, tan clara y libre de artefactos como sea posible.

Sigue una descripción más detallada de la innovación desarrollada para cada uno de éstos módulos.

Creación de base de datos de voz (figura 2)

Este módulo (20) toma las grabaciones de voz originales (21) (grabaciones de la voz de los hablantes de referencia cuya voz quiere imitar el sistema) y las organiza en una base de datos de voz (22), que se usará en el módulo de síntesis. Así que puede decirse que las tareas llevadas a cabo en este módulo tienen el objetivo de

obtener toda la información y las tramas preparadas para usarse directamente en el módulo de síntesis.

Un bloque principal de este módulo tiene la función de incorporar metadatos lingüísticos (añadiendo e indexando metadatos lingüísticos) (23). El otro bloque principal (codificación de voz 24) representa la señal de voz con el modelo de voz usado en el sistema de TTS (tal como se explicó anteriormente, codifica la voz representándola como una secuencia de parámetros). La codificación de voz usada en esta solución es un algoritmo síncrono de periodo fundamental. Esto significa que la tarea es muy sencilla. Está compuesta por un módulo de localización de puntos de inicio (puntos de referencia o "onset points") (25) (tal como se innova mediante la presente invención) y un módulo de codificación de trama (26). El punto crítico en este proceso es la localización de los inicios.

Estos puntos de inicio son puntos de referencia que definen las tramas que deben almacenarse en la base de datos de voz. Es decir, los puntos de inicio marcan el comienzo y el final de cada trama, definiendo por tanto la duración (longitud de trama) y la posición de cada trama. Si la F0 de la trama (F0 local) está definida por la inversa de la longitud de trama, entonces la separación de los puntos de inicio también determina la F0 de la trama. Estas tramas pueden codificarse con cualquier estrategia de codificación de forma de onda (de compresión y expansión como ley A o ley mu, o algoritmos adaptativos tales como modulación por codificación de pulsos diferencial adaptativa ADPCM), con el fin de reducir la huella de la base de datos de voz. Se conserva información de la distancia al inicio previo y la distancia al siguiente.

Estos puntos de inicio definen la separación entre tramas y, como veremos, serán las ubicaciones temporales en las que estarán ubicadas las ventanas para generar las tramas en el módulo de síntesis. Así, la mayor parte del rendimiento del módulo de síntesis dependerá de la correcta ubicación de dichos puntos de inicio.

Tal como se mencionó anteriormente, cuando se describieron los modelos síncronos de pitch de dominio del tiempo pitch de la técnica anterior, necesitan una localización de puntso de inicio muy exacta y constante. El etiquetado manual no es factible para las bases de datos de voz muy extensas usadas hoy en día en sistemas de TTS. Los procesos automáticos aún no ofrecen la precisión y coherencia necesarias para un TTS de alta calidad.

En otros sistemas, como los sistemas de TD_PSOLA, este proceso de localización de los inicios es hace manualmente. La presente invención propone una nueva técnica completamente automática para localizar los puntos de inicio en la señal de voz (llevada a cabo en el módulo de localización de inicios, 25), que es una mejora del modelo de base sinusoidal con análisis mediante método de síntesis (SBMAS). Este método se da a conocer en la patente española P200931212 de Telefónica, que se resume en los siguientes párrafos.

Una de las tareas llevadas a cabo en el modelo SBMAS es la ubicación de ventanas de análisis por medio de un proceso iterativo que calcula la fase del primer componente sinusoidal de la señal de voz y la comparación entre el valor obtenido para esa fase y un valor predeterminado hasta que la ventana de análisis se ubica en un lugar tal que la diferencia de fase hallada en la comparación sea inferior a la mitad de la muestra de voz.

La ubicación de las ventanas de análisis influye en el cálculo de cualquier parámetro estimado a partir de la señal de voz que se ha sometido a enventanado (trama de voz). Las ventanas (que pueden ser de diferente tipo) están diseñadas para enfatizar las propiedades de la señal de voz en su punto central y para atenuarlas hacia sus extremos. El modelo SBMAS se diseñó para mejorar la coherencia en la ubicación de ventanas, para ubicarlas en sitios tan homogéneos como sea posible a lo largo de la señal de voz. Esto se consigue por medio de un proceso de ubicación de ventanas iterativo. En este proceso, para cada ventana seleccionada, se obtiene en las tramas sonoras, cuáles son los parámetros (por ejemplo, la fase) del primer componente sinusoidal de la señal (el más próximo al primer armónico) y comprueba la diferencia entre ese valor y un valor de fase que se toma como referencia (puede considerarse un valor igual a 0 sin perder generalidad). Si la diferencia de fase representa un desplazamiento temporal inferior a un umbral (en una realización preferida, este umbral es la mitad de una muestra de voz), se valida la ubicación de la ventana y la ubicación de la siguiente ventana de análisis comienza tras hacer avanzar la mitad de un periodo fundamental y se repite el proceso para esta nueva ventana de análisis. Si la diferencia de fase representa un desplazamiento temporal igual o superior al umbral, se descarta la ubicación de la ventana y se realiza un nuevo análisis tras mover la ventana un número necesario de muestras de voz y recalcular los parámetros. Este proceso se repite hasta alcanzar la ubicación de ventana correcta (en la que la diferencia entre las fases es inferior al umbral). Una vez alcanzada esta ubicación, la ubicación de la siguiente ventana de análisis comienza tras hacer avanzar la mitad de un periodo fundamental y se repite el proceso. En el caso de hallar una trama sorda durante el proceso, se valida la ubicación de la ventana y la ubicación de la siguiente ventana comienza tras hacerla avanzar 5 ms.

El SBMAS, en el proceso iterativo para optimizar el conjunto de parámetros espectrales para representar cada trama de voz, identifica el conjunto de instantes de sincronía (conocidos en la técnica como "epochs") en los que los componentes espectrales sinusoidales están en fase. Estos instantes "epoch" corresponderán a la ubicación de las ventanas obtenidas en el proceso explicado anteriormente. Estos instantes "epoch" serán los puntos de inicio en una etapa inicial del método propuesto.

Para segmentos con voz (en los que la voz es más periódica) estos instantes "epoch" son muy exactos, y lo

que es más importante, muy constantes por todo el conjunto de estímulos de voz. Cada estímulo de voz es un segmento de voz de la voz grabada del hablante de referencia. El grabador de voz suele ser bastante largo (varias horas) y habitualmente se divide en archivos o segmentos, denominados estímulos de voz.

En segmentos sin voz (por ejemplo en sonidos oclusivos, fricativos, africados), dado que no hay realmente una estructura periódica para hallar un punto en el que los armónicos pudieran estar en fase, los resultados tienen mucho ruido. No es un punto muy crítico, puesto que el TD-PSOLA para estas áreas sordas define una distribución de puntos de inicios en puntos espaciados de manera equidistante (ya que teóricamente no hay necesidad de modificar F0 en segmentos sin voz).

Entonces en los segmentos con voz, la localización de inicios realizada por el sistema SBMAS es bastante exacta de modo que en estos segmentos se conservan los inicios ubicados por el sistema SBMAS.

El problema surge en los segmentos sin voz (también llamados de silencio o sordos) o la transición entre segmentos con voz y sin voz, porque en estas zonas los inicios no son constantes.

Con este objetivo, la siguiente etapa del algoritmo en el módulo propuesto es eliminar por filtración los puntos de inicio no consistentes en los siguientes segmentos:

- segmentos sin voz (en los que tanto la detección de F0 como la identidad de alófonos son de segmentos sin voz)

- segmentos en los que el periodo estimado suavizado (la inversa de la F0 suavizada) y el periodo calculado a partir de la posición de los puntos inicios difiere en más del 50% (estos segmentos serán la transición entre segmentos con voz y sin voz).

Tras este borrado de puntos de inicio, se definen islas de voz con puntos de inicio muy fiables. Normalmente todavía se borran el primer y el último punto de inicio de cada isla si difieren del periodo estimado (facilitado por la F0 suavizada estimada) en más del 20%.

Entonces, las islas de voz se extienden insertando puntos de inicio de manera iterativa, hasta entrar en contacto con otra isla. La estrategia para insertar nuevos puntos de inicio es (para los segmentos citados previamente, en los que se han borrado los puntos de inicio):

En segmentos con voz (o bien por la detección de F0 o por la identidad alofónica) se detectan el mínimo y el máximo de la señal de voz en una porción de longitud la mitad del periodo estimado alrededor del punto inicio siguiente esperado.

- Si el mínimo y el máximo están suficientemente próximos (el 10% del periodo estimado), se selecciona el tiempo de cruce por cero entre el mínimo y el máximo como el nuevo tiempo de inicio.

- Si estos puntos no están suficientemente próximos, entonces se selecciona el cruce por cero más próximo al siguiente inicio estimado.

En segmentos sin voz, se insertan inicios en los tiempos de periodo estimado (inversa de F0 suavizada). Como este periodo procede de la F0 suavizada, que se interpola también para segmentos sin voz, la separación de los inicios evoluciona suavemente de un segmento con voz al siguiente.

Tras llenar un área vacía, si hay una gran discordancia entre la última posición del punto de inicio y el periodo esperado (diferencia mayor que el 20%), la diferencia se redistribuye uniformemente entre los puntos de inicio centrales.

Entonces, tal como puede observarse en la figura 3, en primer lugar se realiza un análisis con base sinusoidal mediante síntesis (SBMAS) y se realiza una primera asignación de puntos de inicios (31). Entonces se filtran (borran) los inicios en segmentos sin voz y segmentos de transición (32) y entonces se colocan los puntos de inicio según el algoritmo anterior (33)

Transformación prosódica

Ya en el sistema de TTS, en el proceso de síntesis del voz, se selecciona el conjunto óptimo de tramas para pronunciar cada alófono del texto de entrada de la base de datos de voz (esto se realiza mediante el módulo de selección de unidad). Estas tramas tienen una F0 original (la F0 local según las grabaciones de voz, determinada en el módulo de creación de base de datos de voz) definida como la inversa del periodo original de las tramas. En este caso, el periodo original de las tramas es la longitud de trama que, tal como se explicó anteriormente, viene dada por la separación de los puntos de inicio.

Para cada alófono, el número de tramas que pertenecen al alófono con su longitud determina la duración

del alófono.

La presente invención también define una nueva estrategia con el fin de solucionar el problema de modificar la F0 y la duración, en el algoritmo de pitch síncronos, minimizando la distorsión acústica.

Esta estrategia funciona de la siguiente manera (véase la figura 4):

La primera etapa (41) consiste en que para un grupo de alófonos entre pausas, cada trama recibe la F0 objetivo según la F0 objetivo de su alófono correspondiente tal como se obtuvo en el módulo de generación prosódica, correlacionando la F0 objetivo del alófono con las tramas que pertenecen al alófono.

Habitualmente, para cada alófono no hay un único valor de la F0 objetivo sino un valor de contorno de la F0 objetivo a lo largo del alófono. Este valor de contorno de F0 se correlaciona con las tramas que pertenecen al alófono, de modo que la F0 objetivo asignada a la trama dependerá de la duración y la situación de la trama en el alófono.

La F0 objetivo asignada a la trama también depende de la estrategia de imposición de F0 (por ejemplo se conserva la F0 original de la trama si difiere en menos del 10% de la F0 objetivo). Como a nivel de trama la F0 y la duración están relacionadas, esta primera etapa ayuda a tener una mejor referencia de cuáles serán los valores finales. Estos valores (temporales) de F0 proporcionarán una primera aproximación al periodo de las tramas, y de ese modo a la duración que tendría la trama si se sintetizara.

Para cada alófono, la estrategia para imponer una duración se describe en el módulo 42. En primer lugar, se impone la duración objetivo (43) para silencio y oclusivas sin voz, ya que van a sintetizarse insertando muestras de amplitud cero. Si no es una oclusiva silenciada y sin voz, se calcula la duración de alófono estimada sumando la duración estimada de las tramas correspondientes (44). Esta duración estimada de cada trama se calcula como la inversa de la F0 objetivo asignada en la etapa previa a cada trama. Dicho de otro modo, para cada alófono, la duración estimada se calculará añadiendo la longitud estimada (la inversa de la F0 objetivo asignada previamente) de las tramas que pertenecen al alófono.

Esta duración estimada calculada para el alófono se compara con la duración objetivo del alófono, producida mediante el módulo generador prosódico (45). Si la diferencia de las duraciones es menor que un umbral, se conserva la duración estimada. Si la diferencia de las duraciones normalizada es mayor que el umbral, la solución para controlar la duración es insertar (si la duración estimada es menor que la duración objetivo) tramas (46) o borrar tramas (47) (si la duración estimada es mayor que la duración objetivo). El umbral depende de la continuidad del alófono:

- Si las tramas en el alófono proceden de dos realizaciones diferentes en los estímulos grabados (segmentos de voz grabada), se define un umbral inferior (15%), dado que ya existe la necesidad de concatenar las muestras, y esto ya producirá cierta distorsión.

- Si todas las tramas en el alófono proceden del mismo alófono original (no hay concatenación de diferentes estímulos), se define un umbral superior (25%), dado que la inserción (o el borrado) de tramas producirá una nueva distorsión que no existía previamente.

La inserción y el borrado de las tramas producirán discontinuidades y distorsión. Con el fin de minimizarlas, la inserción o el borrado de tramas se mantiene en un único punto, el centro del alófono. Finalmente, cuando se insertan nuevas tramas, si las tramas se acaban de replicar, puede haber grandes discrepancias en los contextos que deben solaparse posteriormente para sintetizarse. Debe tenerse en cuenta que los algoritmos TD-PSOLA están orientados principalmente a la modificación de F0; no hay ninguna manera para controlar la evolución espectral en puntos de unión. Con el fin de suavizar este problema, las nuevas tramas se construyen de manera recursiva con la misma técnica de solapamiento y adición usada en la síntesis final de voz. Este proceso se explicará con más detalle en la siguiente sección (y puede observarse en la figura 7, cuando se presente el algoritmo básico):

- En primer lugar se establecen el punto de inserción y las tramas adyacentes. Si las tramas en la secuencia original proceden de segmentos de voz no consecutivos (si es un “punto de adhesión” en el que deben unirse tramas de dos segmentos de voz diferentes), entonces se selecciona este límite entre las tramas de una unidad y la otra. Como ya habrá una discontinuidad espectral en este punto, es el mejor punto para mantener estas discontinuidades a un mínimo. Si no hay ningún “punto de adhesión” (si la secuencia original de tramas procede de un único segmento de voz), entonces se selecciona la trama central como punto de inserción, y se usan las tramas previa y siguiente como fuentes para construir las nuevas tramas de solapamiento y adición. Esta trama central se construye por el algoritmo de solapamiento y adición modificado entre las tramas a ambos lados del punto de inserción. En el ejemplo mostrado en la figura 7, el punto de inserción será el punto entre las tramas B y C. Si deben insertarse más tramas, el proceso se repite de manera iterativa en cada mitad, tomando las tramas recién generadas como “donantes”. Es decir, en cada etapa, si el número de tramas que debe crearse es una, entonces es sólo el resultado de aplicar el modelo modificado de solapamiento y adición a las tramas “donantes” tal como se explicó (o bien tramas originales a los lados del punto de inserción, o bien

tramas sintetizadas en una etapa previa de este proceso); si el número de tramas que debe crearse es de dos, la primera se crea tal como se describió, y la siguiente se crea mediante el solapamiento y la adición entre ésta y el "donante" a la derecha; si deben crearse más tramas, entonces la trama central se crea tal como se describió y se repite el proceso para el resto de las tramas en dos bloques, una mitad entre el "donante" izquierdo y la trama recién creada, y el resto entre esta trama y el "donante" a la derecha.

Esta estrategia produce una transición suave en las tramas insertadas y mantiene la distorsión producida por los contextos alejados a un mínimo.

Para cada trama, la estrategia para imponer una F0 se describe en el módulo 48. Tras modificar (si fuera necesario) el número de tramas, la F0 objetivo para cada trama vuelve a calcularse (49) (ya que la correlación de la F0 objetivo del generador prosódico depende de la duración real y la distribución de tramas en el alófono). Entonces, la F0 objetivo se impone o no dependiendo de su diferencia con respecto a la F0 original. Si la diferencia es menor que un umbral (492), se conserva la F0 original, si no, se impone la F0 objetivo (493). De nuevo, el umbral depende de la concatenación o no en el alófono (10% si hay concatenación, 20% si no). Como diferencia se suele usar la diferencia relativa, es decir, la diferencia absoluta partida por el valor de F0 objetivo

Si la secuencia de tramas no es original (es decir, se han añadido en el proceso de ajustar la duración del alófono tal como se explicó anteriormente), hay un ajuste de amplitud (494).

Si el alófono no es silencio o una oclusiva sin voz, y el siguiente alófono no es un silencio o una oclusiva sin voz, se calcula la amplitud pico en las tramas de borde de cada alófono (495), amplitud si la diferencia de la amplitud pico es mayor que un umbral (por ejemplo el 25%) (496), hay una última interpolación de amplitud (497). Para esta interpolación, pueden usarse por ejemplo las siguientes fórmulas de ecualización:

$$F_{r1}=(k*Max_1+(1-k)*Max_2)/Max_1$$

$$F_{r2}=(k*Max_2+(1-k)*Max_1)/Max_2$$

en las que k es un parámetro de diseño, Max_1 y Max_2 son la amplitud pico para la primera y la segunda trama respectivamente y F_{r1} y F_{r2} son los factores de multiplicación para suavizar la amplitud para la primera y la segunda trama respectivamente.

Si la diferencia de la amplitud pico es menor que un umbral (por ejemplo el 25%) (498), no se realiza ninguna interpolación de la amplitud.

Síntesis de señal de voz

Lo básico de los algoritmos de solapamiento y adición síncrono con el periodo fundamental de la técnica anterior se muestra en las figuras 5 y 6.

El método de la técnica anterior para sintetizar tramas de voz será tal como sigue:

En primer lugar las tramas originales se enventanan. Las ventanas se centran en los puntos (instantes) de inicio; en este caso los puntos de inicio definirán la separación de las tramas que van a tratarse.

Las ventanas (habitualmente una ventana de Hamming, pero que puede ser de un tipo diferente) están diseñadas para enfatizar las propiedades de la señal de voz en su centro, y están atenuadas en sus extremos. La función de la ventana es suavizar el solapamiento y la adición cuando se modifica la F0, y también suavizar la concatenación de tramas que proceden de diferentes segmentos. La ventana de Hamming (y las otras, como la de Hanning) es simétrica, está centrada en el tiempo de inicio y pasa al inicio previo (o siguiente).

En la figura 5, A_{LW}, B_{LW}, C_{LW} representan las tramas originales A, B y C respectivamente que se han enventanado a la izquierda (es decir, multiplicadas por el ala izquierda de la ventana centrada en la separación entre las tramas A y B, entre las tramas B y C y entre las tramas C y D respectivamente). B_{RW}, C_{RW}, D_{RW} representan las tramas originales B, C y D respectivamente que se han enventanado a la derecha (es decir, multiplicadas por el ala derecha de la ventana centrada en la separación entre las tramas A y B, entre las tramas B y C y entre las tramas C y D respectivamente).

Tras enventanar las tramas, las tramas se "separan" (es decir, se extienden o se reducen) según el nuevo periodo objetivo, el inverso de la F0 objetivo de cada trama (véase la figura 6).

La "separación" de las tramas según el nuevo periodo objetivo se realiza de cualquiera de las maneras bien conocidas y es una adaptación de la duración de las tramas al nuevo periodo objetivo. Por ejemplo, puede realizarse añadiendo muestras con valor cero en el extremo de la trama, si el periodo objetivo es mayor que el periodo original, y borrando algunas de las muestras del extremo de la trama, si el periodo objetivo es menor que el periodo original.

Con el fin de tener una imagen más clara, en la figura 5 sólo se muestran las etapas de un algoritmo de

solapamiento y adición puro, no mostrándose explícitamente esta etapa “de separación” intermedia.

Tras esta separación, se obtienen después tramas solapando y añadiendo las tramas tal como se muestra en la figura 5, es decir añadiendo las muestras de B_{RW} con B_{LW} y de C_{RW} con C_{LW}.

En el algoritmo de la técnica anterior, la ventana es simétrica y tiene una longitud L que es el doble del periodo original de la señal de voz T_0 (véase la figura 6, primer gráfico). En la figura 6, segundo gráfico, se muestra cómo las muestras se separan según el periodo objetivo T y entonces se añaden las muestras (tercer gráfico).

En el algoritmo de la técnica anterior, el enventanado se realiza según la duración de trama original en el proceso de construcción de base de datos (de modo que se ahorra tiempo de procesamiento cuando se sintetiza).

Usando la duración original (T_0) en el enventanado, tiende a producir “huecos” en el centro de la trama sintetizada cuando la duración es mayor que la original (disminución de la F0) y “saltos” cuando se acortan las tramas (aumento de la F0).

En la solución propuesta, el enventanado se lleva a cabo en el momento de síntesis de voz, de modo que en las ventanas puede usarse el periodo objetivo de la trama, en lugar del original. Esta solución justo a tiempo soluciona el problema presentado previamente, que se nota más cuando están implicadas grandes modificaciones de F0.

La otra innovación propuesta para este módulo es el uso de ventanas asimétricas.

Cuando se manejan no sólo grandes modificaciones prosódicas, sino también movimientos de F0 rápidos, la F0 objetivo (y por consiguiente la duración de trama o periodo objetivo) puede ser diferente de una trama a la siguiente. En el algoritmo original, la forma de la ventana es simétrica, de modo que se asume el mismo periodo para ambos lados del inicio (por ejemplo para dos tramas secuenciales). Cuando existen movimientos de F0 rápidos, este efecto produce distorsión, incluso si no hay modificación prosódica.

En la solución propuesta, la partición en ventanas se calcula independientemente para ambas tramas consecutivas (cola derecha e izquierda para el inicio). Esto reduce la cantidad de ruido generado y produce una mejor calidad cuando se manejan movimientos de F0 rápidos. Es decir, si la F0 objetivo para la trama A es $F0_A$ (el periodo objetivo de trama será $1/F0_A$) y la F0 objetivo para la trama B es $F0_B$ (el periodo objetivo de trama será $1/F0_B$), entonces la ventana usada será de una duración $1/F0_A$ para el ala izquierda y $1/ F0_B$ para el ala derecha.

Tras esta partición en ventanas asimétrica, se aplica el método clásico. Es decir, las tramas se “separan” (se extienden o se reducen) según el nuevo periodo objetivo y entonces las tramas se solapan y se añaden (tal como se muestra en la figura 5).

Finalmente, ahora puede explicarse más detalladamente la innovación en la generación de tramas presentada en la sección previa. Tal como ya se indicó, los algoritmos TD-PSOLA se centran en modificaciones de F0, y la partición en ventanas y la adición no son herramientas muy potentes para combinar diferentes tramas (ya que supone que la voz es muy estacionaria y periódica). Esto se aplica también cuando se añaden nuevas tramas en el proceso de ajuste de la duración (tal como se observó por ejemplo anteriormente en el módulo de transformación prosódica).

En los métodos de la técnica anterior, esta adición de tramas se realiza simplemente replicando las tramas existentes. Pero replicar una trama produce la necesidad de añadir a la ventana alas que proceden de diferentes tramas. Esto producirá una distorsión, más importante si el estacionario es cuestionable, tanto de manera acústica como de manera prosódica. En otras soluciones replicar la trama comprende construir partes de tramas artificialmente que pueden producir discontinuidades en el punto de unión.

Por tanto la solución propuesta es intentar crear tramas “sintéticas” en lugar de replicadas. Para esto, se generan nuevas tramas a partir de las tramas adyacentes que se han sometido a enventanado y entonces estas tramas generadas se añaden a la secuencia de tramas con un algoritmo de solapamiento y adición clásico.

La nueva estrategia se presenta en la figura 7.

En este algoritmo, se enventanan (se someten a partición en ventanas) tres tramas consecutivas (A, B y C) como en el procedimiento normal. Entonces se construyen dos tramas generadas nuevas; una de ellas se obtiene añadiendo las muestras de las partes que se han enventanado a la izquierda de dos tramas consecutivas (A_{LW} + B_{LW}) y la segunda se obtiene añadiendo las partes que se han enventanado a la derecha de las dos tramas consecutivas siguientes (B_{RW} + C_{RW}). Todas estas partes se producen cuando se enventanan las tramas originales de modo que no hay ninguna creación artificial de ninguna parte.

Entonces estas tramas generadas se insertan en un proceso de solapamiento y adición normal, en el punto de inserción entre las tramas B y C. Es decir, la primera trama generada se añade a la parte que se ha enventanado a la derecha de la segunda trama (B_{RW}) para obtener una nueva trama y la segunda trama generada se añade a la

parte que se ha enventanado a la izquierda de la segunda trama (B_{LW}), obteniendo dos nuevas tramas B_{RW}+ (A_{LW} + B_{LW}) y (B_{RW} + C_{RW}) + B_{LW} que remplazarán la trama B.

Con esta nueva generación de tramas se alcanza una evolución más suave que ha demostrado proporcionar una mejor calidad que la replicación simple de tramas.

Resumiendo, la solución propuesta presenta un nuevo método para la síntesis de voz que soluciona alguno de los inconvenientes de los métodos previos.

El método de SBMAS (dado a conocer en la patente española P200931212 “Codificación, Modificación y Síntesis de Segmentos de Voz”) ofrece una alta calidad, suavidad y flexibilidad para grandes modificaciones prosódicas, pero presenta dos problemas importantes cuando se usa en dispositivos no muy potentes, como teléfonos móviles: 1) consume mucha CPU; 2) la base de datos de voz es muy grande. No sólo los requisitos de espacio del disco, sino también los requisitos de memoria cuando se sintetiza. Estos factores no son tan relevantes en aplicaciones de escritorio o soluciones de IVR de plataforma grande, pero pueden ser cruciales en dispositivos de baja capacidad.

La presente invención permite el uso de modelos de codificación muy simples y eficaces, desde el punto de vista de la memoria, en comparación con el modelo definido por el SBMAS. Por ejemplo, para una frecuencia de muestra de 8 KHz, usar una PCM de 16 bits necesitaría un uso de memoria de 128 Kbits por segundo de voz, en comparación con los 300 Kbits (variables) del SBMAS. Este requisito puede reducirse adicionalmente con el uso codificación mediante ley A (64 Kbits) o incluso ADPCM (32 Kbits).

Con el fin de solucionar estos problemas, se ha desarrollado un módulo mucho más ligero, basándose en TD-PSOLA (algoritmo de solapamiento y adición de pitch síncrono en el dominio del tiempo), sin modelo de voz explícito (sólo las muestras, o una forma de onda codificada de las mismas). Por tanto, la base de datos de voz puede ser de la misma magnitud de los estímulos grabados o incluso la mitad o un cuarto de su tamaño, si se usa la compresión y expansión mediante ley A o codificación mediante ADPCM (como referencia, el SBMAS puede requerir una base de datos de voz que sea dos veces los estímulos grabados) y, para el tiempo de procesamiento de la CPU, la nueva solución sólo emplea básicamente dos multiplicaciones y una suma por muestra, mucho menos que el SBMAS que requiere operaciones complejas como arco tangente y exponenciales.

Con respecto al TD-PSOLA convencional, el nuevo método usa muchas soluciones para mejorar la calidad, principalmente cuando se manejan modificaciones prosódicas grandes y voz expresiva:

- La duración se modifica tras imponer la F0 objetivo, y sólo si la diferencia entre la duración original y la duración objetivo es mayor que un umbral que depende de si ha habido concatenación en el centro del alófono, o es uno continuo.

- La duración se transforma insertando o borrando tramas. Pero las tramas insertadas no son sólo una réplica de las de los bordes, sino que se construyen de manera recursiva con la misma estrategia de solapamiento y adición.

- La F0 objetivo sólo se impone si su diferencia con respecto a la original es mayor que un umbral. Este umbral depende de si ha habido una concatenación en el centro del alófono, o es uno continuo.

- En la síntesis, las tramas se enventanan justo a tiempo (“just in time”) con el periodo objetivo en lugar de realizarse en la fase de construcción de base de datos con el periodo original. Esto reduce la distorsión cuando se manejan modificaciones de F0 grandes.

- Las ventanas usadas son asimétricas, permitiendo diferentes longitudes para las mitades derecha e izquierda de las tramas. Esto ofrece una mejor calidad cuando se manejan movimientos de F0 muy rápidos (como en la voz expresiva).

Estas soluciones permiten el uso de un algoritmo similar a TD-PSOLA en la síntesis de un voz más expresiva que la que se ha empleado habitualmente.

Aunque la presente invención se ha descrito con referencia a realizaciones específicas, los expertos en la técnica deben entender que los anteriores y diversos otros cambios, omisiones y adiciones en la forma y el detalle de las mismas pueden realizarse sin apartarse del espíritu y del alcance de la invención tal como se definen mediante las siguientes reivindicaciones.

REIVINDICACIONES

1. Un método para la síntesis de señal de voz, en el que cada alófono que va a reproducirse en la señal de voz sintetizada tiene un valor objetivo deseado de duración y un valor objetivo deseado de frecuencia fundamental, denominado F0 objetivo, y en el que la señal de voz que va a sintetizarse se aparta de las unidades de señal de voz grabadas previamente de un hablante de referencia, estando cada unidad de señal de voz compuesta por una secuencia de tramas de señal de voz, denominadas tramas originales, teniendo cada trama original una frecuencia fundamental F0, denominada F0 original, y en la que, dada la secuencia de alófonos que va a reproducirse, se selecciona una secuencia de tramas de señal de voz originales correspondiente a dicha secuencia de alófonos, comprendiendo el método las siguientes etapas:

a) Asignar una F0 objetivo a cada una de las tramas originales de la secuencia seleccionada de tramas originales, basada en la F0 objetivo del alófono correspondiente, siendo el periodo objetivo asignado a cada trama 1/F0 objetivo de la trama.

b) Generar la señal de voz, comprendiendo esta etapa:

b1) Modificar la secuencia de tramas originales, enventanando dicha secuencia de tramas originales estando las ventanas centradas en el punto de separación entre cada dos tramas consecutivas, siendo las ventanas asimétricas, calculándose la longitud de la ventana de manera independiente para ambas tramas consecutivas situada cada una a un lado del punto donde se centra la ventana, es decir, siendo la longitud del ala derecha de la ventana el periodo objetivo de la trama situada a la derecha del punto en el que se centra la ventana y siendo la longitud del ala izquierda de la ventana, el periodo objetivo de la trama situada a la izquierda del punto en el que se centra la ventana.

2. Un método según la reivindicación 1, en el que la separación entre tramas consecutivas de la secuencia seleccionada de tramas originales viene dada por la ubicación de puntos de tiempos de referencia denominados puntos de inicio que definirán el final de una trama y el principio de la siguiente trama y por consiguiente, la ubicación de los puntos de inicio determinará la longitud de cada trama y en el que la F0 original se calcula como la inversa de la longitud de cada trama

3. Un método según la reivindicación 1, en el que la etapa de asignación de una F0 objetivo a cada una de las tramas originales de la secuencia basada en la F0 objetivo del alófono correspondiente, comprende:

- Calcular una F0 objetivo inicial para cada una de las tramas originales de la secuencia según la F0 objetivo del alófono correspondiente y si la diferencia entre la F0 original y la F0 objetivo calculada es mayor que un primer umbral, dicha F0 objetivo calculada se asigna como la F0 objetivo para la trama y si no, la F0 original de la trama se asigna como la F0 objetivo para la trama;

- Para cada alófono que va a reproducirse, que no es un alófono oclusivo sin voz o en silencio, la duración estimada de alófono se calcula añadiendo la duración estimada de las correspondientes tramas que forman el alófono, siendo la duración estimada de cada trama la inversa de la F0 objetivo asignada a la trama en la etapa anterior; entonces esta duración de estimación se compara con la duración objetivo de alófono y si la diferencia es menor que un segundo umbral, la duración estimada se mantiene (46) y si la diferencia de las duraciones es mayor que el segundo umbral, entonces se cambia la duración del alófono insertando tramas (46) si la duración estimada es menor que la duración objetivo, o borrando tramas (47) si la duración estimada es mayor que la duración objetivo;

- Como las tramas de cada alófono pueden haber cambiado en la etapa anterior, para cada trama, se calcula de nuevo la F0 objetivo según la F0 objetivo del alófono correspondiente y si la diferencia entre la F0 original y la F0 objetivo recalculada es mayor que un tercer umbral, se asigna dicha F0 objetivo calculada como la F0 objetivo para la trama y si no, se asigna la F0 original de la trama como la F0 objetivo para la trama.

4. Un método según la reivindicación 3, en el que para cada alófono, el valor de los umbrales segundo y tercero depende de si las tramas asignadas al alófono proceden de diferentes realizaciones de voz originales del alófono del hablante de referencia o de la misma realización de voz original del alófono.

5. Un método según la reivindicación 4, en el que el umbral segundo y tercero tienen un valor del 15% para los alófonos cuyas tramas proceden de dos diferentes realizaciones de voz originales del alófono y un valor del 25% para los alófonos cuyas tramas proceden de la misma realización de voz original del alófono.

6. Un método según la reivindicación 3, en el que la etapa de adición de tramas al alófono, si la duración estimada es menor que la duración objetivo, se realiza generando nuevas tramas usando las tramas enventanadas adyacentes al punto en el que desea añadirse una nueva trama y entonces estas tramas generadas se añaden a la secuencia de tramas enventanadas con un algoritmo de solapamiento y adición.

7. Un método según la reivindicación 2, que incluye además previo a la etapa a), una etapa de ubicación de

los tiempos de inicio en la secuencia de tramas originales, que incluye las siguientes acciones:

- Identificar los tiempos en los que los componentes espectrales sinusoidales de la secuencia de tramas originales están en fase y se toman estos tiempos como la ubicación inicial de los puntos de inicio;

- Obtener una primera estimación de la frecuencia fundamental de cada trama y obtener una primera estimación del periodo fundamental como la inversa de la primera frecuencia fundamental estimada;

- En segmentos sin voz de la secuencia de tramas originales, la separación de los puntos de inicio se dará mediante el periodo fundamental estimado;

- En segmentos con voz en los que la diferencia entre la separación de los puntos de inicio al principio y el primer periodo fundamental estimado está por encima de un cierto cuarto umbral, se detectan el mínimo y el máximo de la señal de voz en una porción de longitud la mitad del periodo estimado alrededor del punto inicio siguiente esperado y:

  - Si la distancia entre el mínimo y el máximo está por debajo de un quinto umbral, se selecciona el tiempo de cruce por cero entre el mínimo y máximo como un nuevo tiempo de inicio;

  - Si la distancia entre el mínimo y el máximo está por encima del quinto umbral, entonces se selecciona el cruce por cero más cercano al punto inicio siguiente estimado como un nuevo tiempo de inicio.

- En segmentos con voz en los que la diferencia entre la separación de la ubicación inicial de los puntos de inicio y el periodo estimado está por debajo del cuarto umbral, la ubicación inicial de puntos de inicio se toma como la ubicación definitiva de los puntos de inicio.

8. Un método según la reivindicación 7, en el que la primera estimación de la frecuencia fundamental se realiza usando un algoritmo de detección de periodo fundamental.

9. Un método según la reivindicación 7, en el que el cuarto umbral es el 50% y el quinto umbral es el 10%.

10. Un método según la reivindicación 1, en el que la etapa de generación de las señales de voz incluye además:

b2) tras enventanar las tramas, ajustar la separación de las tramas según su periodo objetivo;

b3) añadir las tramas ajustadas enventanadas usando un algoritmo de solapamiento y adición.

11. Un método según la reivindicación 1, en el que la F0 objetivo para el alófono es una F0 de contorno objetivo compuesta por tres valores, en el principio, en el centro y en el final del alófono.

12. Un sistema que comprende medios adaptados para llevar a cabo el método según cualquier reivindicación anterior.

13. Un programa informático que comprende medios de código de programa informático adaptados para llevar a cabo el método según cualquiera de las reivindicaciones 1 a 11 cuando dicho programa se ejecuta en un ordenador, un procesador de señal digital, un disposición de puertas programables en campo, un circuito integrado de aplicación específica, un microprocesador, un microcontrolador, o cualquier otra forma de hardware programable.

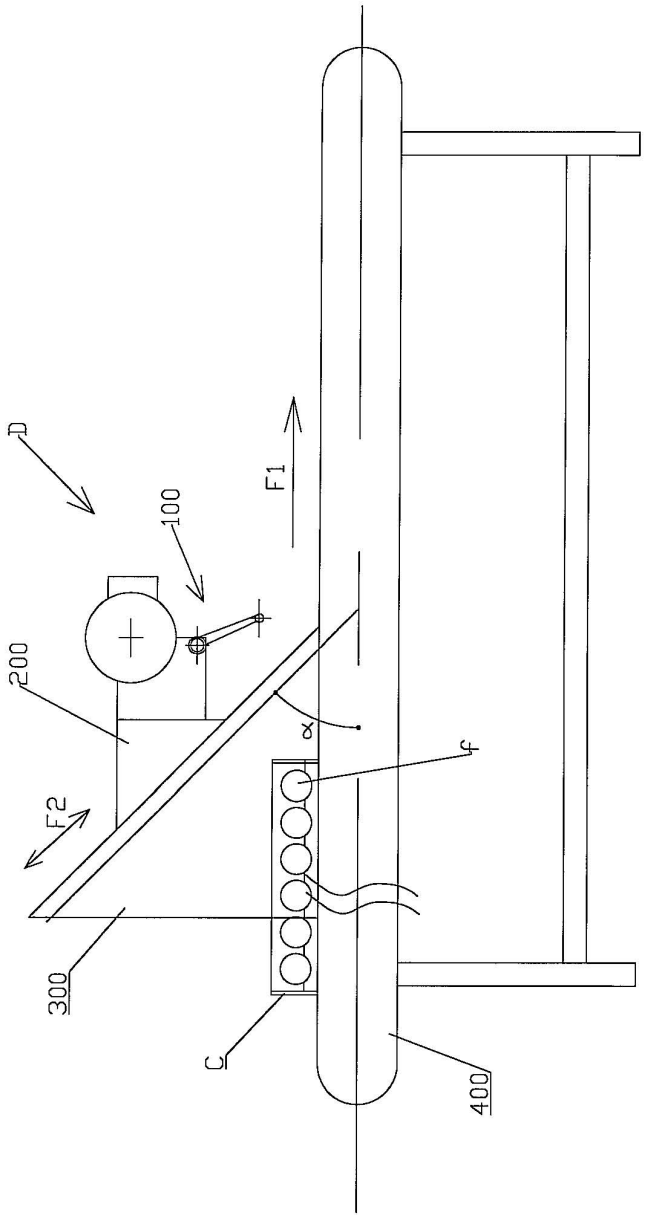


Fig. 1

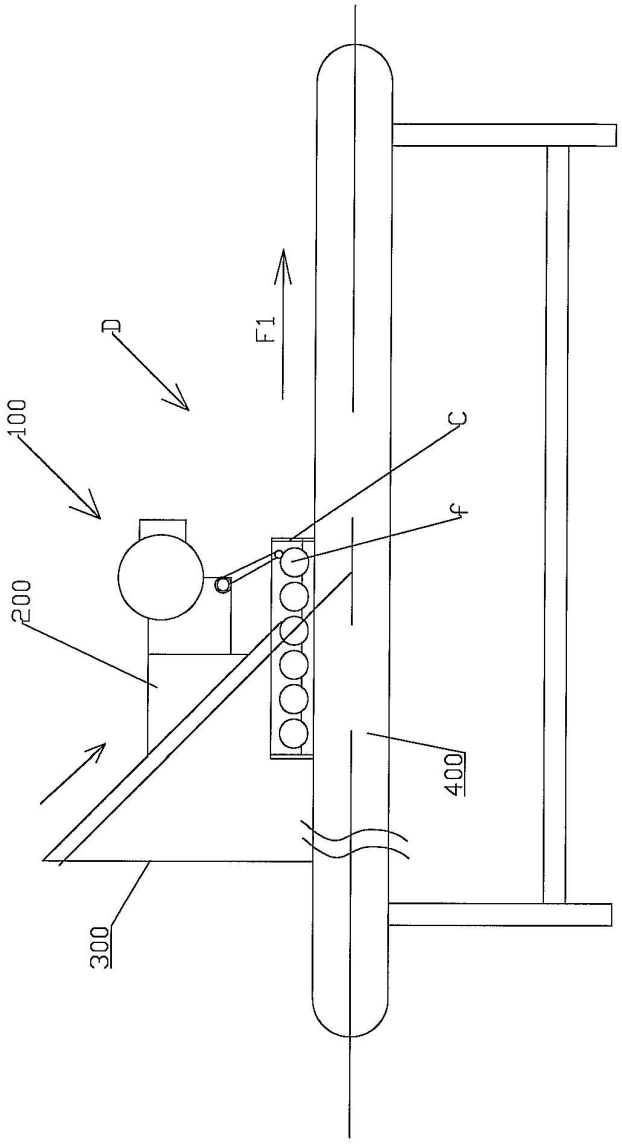
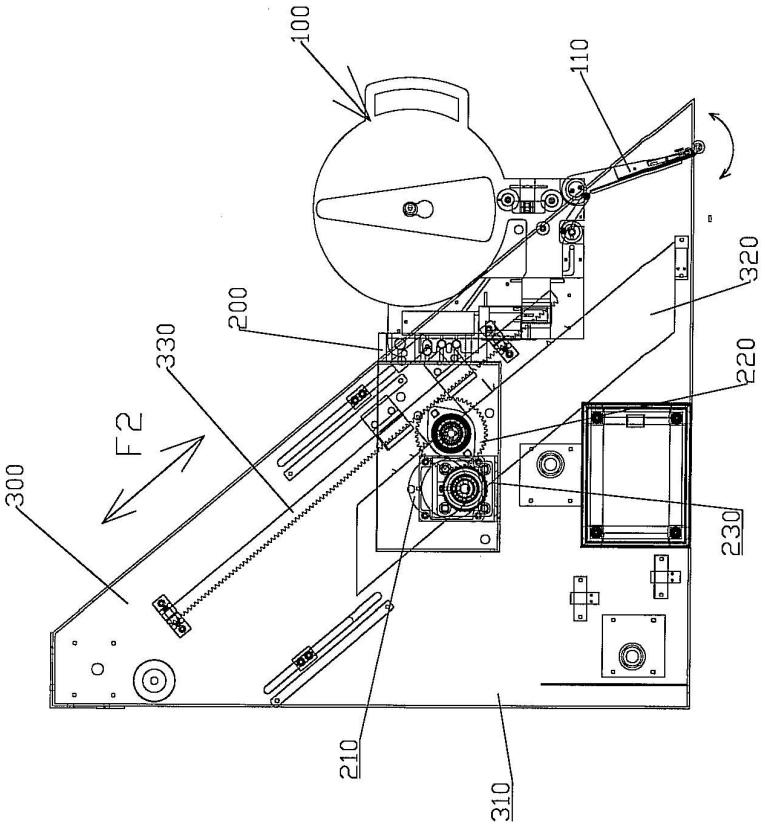


Fig. 2

Fig. 3



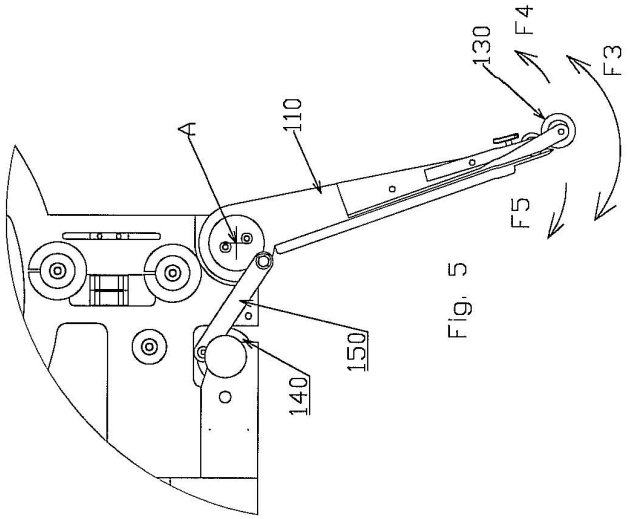


Fig. 5

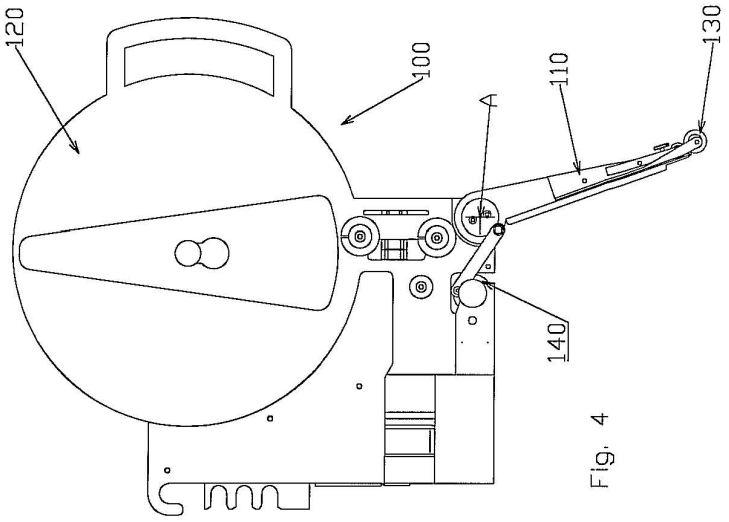


Fig. 4