(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 382 319**

(21) Número de solicitud: 201000256

(51) Int. Cl.:

G10L 13/02 (2006.01)

G10L 13/08 (2006.01)

G10L 13/06 (2006.01)

(12) SOLICITUD DE PATENTE A1

(22) Fecha de presentación: **23.02.2010**

(43) Fecha de publicación de la solicitud: **07.06.2012**

(43) Fecha de publicación del folleto de la solicitud: **07.06.2012**

(71) Solicitante/s: **Universitat Politècnica de Catalunya**
c/ Jordi Girona, 31
08034 Barcelona, ES

(72) Inventor/es: **Álvarez Flórez, Jesús Andrés;**
Vilà Fumàs, Pere;
Moreno Eguilaz, Manuel;
Miserachs Teixidor, Jordi;
Sánchez Alvira, Jordi;
Aymerich Capdevila, Nivard;
Armisen Morell, Albert y
Musquera Moreno, Marc

(74) Agente/Representante:
No consta

(54) Título: **Procedimiento para la síntesis de difonemas y/o polifonemas a partir de la estructura frecuencial real de los fonemas constituyentes.**

(57) Resumen:
Procedimiento para la síntesis de difonemas y/o polifonemas a partir de la estructura frecuencial real de los fonemas constituyentes.
La presente invención se refiere al procedimiento para la generación de la señal acústica de voz sintética de sonidos a partir de una mínima información previa de los fonemas constituyentes, obtenidos por segmentación de una grabación previa; dicha información mínima consiste en la envolvente espectral correspondiente a un único periodo situado en la zona más estacionaria del fonema.

ES 2 382 319 A1

DESCRIPCIÓN

Procedimiento para la síntesis de difonemas y/o polifonemas a partir de la estructura frecuencial real de los fonemas constituyentes.

Sector de la técnica

Sistemas de síntesis de voz.

Antecedentes de la invención

Los sistemas de síntesis de voz actuales se basan en la concatenación de segmentos de voz natural para lo que es necesario un conjunto de palabras grabadas previamente, denominado corpus. Este corpus ha de ser lo suficientemente grande ya que de él se obtendrán los segmentos de voz a concatenar dependiendo de: su posición relativa, características entonativas y duración. La voz ha de ser segmentada a múltiples niveles: semifonemas, fonemas, difonemas, trifonemas, sílabas y hasta palabras o grupos de palabras.

Los sistemas de síntesis de voz actuales basados en la concatenación de segmentos de voz realizan el proceso de síntesis en las siguientes etapas:

a) Selección de unidades previamente segmentadas.

b) Modificación y ajuste de sus características suprasegmentales para la adaptación prosódica al nuevo contexto.

c) Concatenación de los segmentos de voz mediante suma en el dominio temporal.

Objeto y Resumen de la invención

La presente invención pretende resolver el problema de la generación de la señal acústica de voz sintética sin la necesidad de disponer de un gran corpus, por tanto, con una mejora desde el punto de vista económico y del tiempo de elaboración.

De acuerdo con este objetivo la presente invención se refiere al proceso de generación de señales que reproduzca el tránsito entre estados estacionarios de los fonemas a generar.

La transición entre estados estacionarios de los fonemas a generar es posible debido a que la señal de voz natural está compuesta por una sucesión de estados cuasi estacionarios correspondientes a los fonemas que la componen, y a las transiciones continuas de unos fonemas a otros.

Los fonemas sobre los que se transita para la obtención de señales han de ser obtenidos en una fase previa en la que voces grabadas son segmentadas para la obtención de los fonemas constituyentes.

El proceso de tránsito entre fonemas descrito permite la preasignación de las evolventes de las características prosódicas implícitas en la señal portadora generada, como entonación, intensidad y duración de los fonemas; necesario para la transmisión de un mensaje emocional que se acerque en calidad al habla natural.

El procedimiento para el tránsito permite, además de lo expuesto en el párrafo anterior, evolucionar desde la composición frecuencial del fonema de origen al de destino por caminos en los que en ningún momento la percepción sonora es disonante.

Para poder cumplir con las características mencionadas, durante el tránsito se ha de poder modificar la frecuencia fundamental y la energía de los fonemas constituyentes.

También se ha de producir esta transición sin que durante la misma se produzcan discontinuidades audibles.

En un primer paso se procede a la obtención de los coeficientes de la serie de Fourier de los fonemas sobre los que se aplicará el tránsito.

Cada uno de los fonemas queda caracterizado por una serie de armónicos. Cada armónico es un tono puro en fase, frecuencia y amplitud.

De acuerdo con la realización elegida, el paso entre fonemas se lleva a cabo mediante la transición continua entre parejas de armónicos del mismo orden del estado estacionario del fonema inicial y el estado estacionario del fonema final.

El estado inicial respecto al final, sobre los que se realiza el tránsito, puede contener diferente número de armónicos. Para ello se incorporan armónicos de amplitud cero al estado que se encuentre en minoría hasta completar las parejas de armónicos entre el estado inicial y el final.

Se establecen los puntos temporales de tránsito, tiempo inicial y final, y por tanto su duración.

Con el fin de evitar discontinuidades en la señal generada se ha de imponer que la transición entre el estado inicial y final tenga continuidad en amplitud, continuidad en fase instantánea y continuidad en frecuencia instantánea.

En la transición de cada pareja de armónicos, del fonema inicial y del fonema final, se fuerza una transición de fases en tres tramos.

Un primer tramo lineal para valores temporales inferiores al tiempo inicial de tránsito, cuyo valor de fase corresponde a los componentes del estado inicial.

Un segundo tramo cuadrático para valores temporales comprendidos entre el tiempo inicial y final de tránsito.

El segundo tramo ha de ser cuadrático para asegurar la continuidad de la frecuencia instantánea en el inicio y final de la transición.

Un tercer tramo lineal para valores temporales superiores al tiempo final de tránsito, cuyo valor de fase corresponde a los componentes del estado final.

La transición de frecuencias y fases se puede llevar a cabo componente a componente estableciendo una función de tránsito de frecuencias e imponiendo el valor de fase instantánea al inicio de la transición así como al final.

Preferiblemente, la frecuencia fundamental de la señal de transición se debe situar entre las frecuencias fundamentales de la señal inicial y final; así se evita la generación de ruidos debidos al aumento y después a la disminución (o viceversa), en un breve espacio de tiempo, de la frecuencia de la señal.

En este caso, debido a que la pendiente de la fase corresponde a la frecuencia instantánea, el valor de esta pendiente ha de situarse entre los valores de pendiente de fase del estado inicial y final.

Dependiendo de la evolución de las fases del estado inicial y final, la fase de la señal de transición puede tener una pendiente mayor, menor o situarse en un valor intermedio de la pendiente del estado inicial y final.

En algunas realizaciones, para evitar la obtención de una señal de transición con una frecuencia superior o inferior a las frecuencias de los estados inicial y final se realiza una corrección sumándole o restándole una fase llamada α a la fase del componente del fonema inicial o final.

El valor de esta fase α provoca un retardo o un adelanto en el tiempo de la componente a la que se le ha aplicado la corrección de fase α.

Con el fin de que el efecto de la corrección α afecte a todo el fonema, a cada componente de fase del fonema se le aplica la corrección de fase α.

Para minimizar el recorrido de corrección, la obtención del valor de la fase a se inicia con la corrección previa de $\pm 2\pi$ radianes a los componentes de fase del fonema a los que se le suma o resta la fase α.

La fase α es un valor de compromiso de los diferentes componentes del fonema: componentes de fase o componentes de fase y amplitud, en el que se tiene en cuenta tanto el fonema inicial como el final.

Otro aspecto a considerar en el tránsito de los estados inicial y final es la función que sigue la transición.

Esta función de transición ha de proporcionar unos resultados sintéticos que se ajusten al patrón de voz real.

De acuerdo con la realización preferida, el procedimiento de transición, en su conjunto, depende de distintos factores para aplicar de forma específica el tránsito entre los estados inicial y final:

a) Camino de fases a seguir para convertir el fonema 1 en el 2.

b) Función de tránsito.

c) Punto de tránsito.

d) Duración del tránsito.

Breve descripción de los dibujos

Para mayor compresión de cuanto se ha expuesto se acompañan unos dibujos sólo a título de ejemplo no limitativo.

En los dibujos:

La figura 1 es una gráfica que muestra el espectro frecuencial y su envolvente, en módulo, correspondiente al fonema "m" de la palabra "anomena"; El eje de abscisas tiene unidades de frecuencia (Hz); El eje de ordenadas muestra el módulo en escala logarítmica;

La figura 2 es una gráfica que muestra la señal temporal de la palabra "anomena"; El eje de abscisas tiene unidades temporales (ms); El eje de ordenadas muestra la amplitud normalizada;

La figura 3 es una gráfica que muestra una zona ampliada de la figura 2; Por tanto, sus ejes tienen las mismas unidades, respectivamente, que la figura 2; y

La figura 4 es una gráfica que muestra la transición de la fase y de la frecuencia; El eje de abscisas tiene unidades temporales (s); El eje de ordenadas tiene unidades de ángulo plano (rad).

Descripción de realizaciones preferidas

En la figura 1 se muestra la transformada rápida de Fourier a partir de la señal temporal de un fonema de voz real, en este caso el fonema "m".

La componente 1 corresponde al módulo de la señal para una frecuencia dada.

La evolvente espectral 2 se obtiene a partir de los espectros frecuenciales que componen la señal discretizada.

La figura 2 muestra la señal temporal de voz real de la palabra "anomena". El tramo de señal 3 es el correspondiente a la transición del fonema "o" al fonema "m".

La figura 3 muestra con más detalle el tramo de señal 3. Con este detalle se aprecia cómo la voz natural está compuesta por estados cuasi estacionarios.

El período 4 es el período de transición desde el fonema "o" al fonema "m".

La figura 4 muestra las fases y las frecuencias instantáneas del estado inicial y final para una pareja de armónicos, así como sus caminos de evolución o transito.

Se establece el punto temporal de inicio de la transición 5 y el de fin 6.

Por tanto, quedan determinados los tres tramos de la transición. Primer tramo para valores temporales inferiores al punto temporal de inicio de transición 5. Segundo tramo comprendido entre el punto de inicio de transición 5 y el punto de fin de transición 6. Y un tercer tramo para valores temporales superiores al punto temporal 6.

A la fase instantánea del estado final 7 se le suma una fase α 8. De esta suma de fases surge la fase instantánea 9.

En el caso mostrado, el tránsito entre fases se realiza entre la fase instantánea del estado inicial 10 y la fase instantánea 9.

La mejora introducida con la adhesión de la fase α 8, en este ejemplo, se obtiene en el hecho de que el valor de la frecuencia instantánea de transición 11, es superior a la frecuencia instantánea del estado inicial 12 e inferior a la frecuencia instantánea del estado final 13; y por tanto, no se produce un altibajo brusco de la frecuencia de la señal con el consecuente ruido generado.

Es evidente que la figura 4 muestra un caso particular de transito y por tanto la fase α podría restarse así como no estar presente, según los casos.

Según el planteamiento mostrado, también sería posible que la fase α se sumarse o restarse a la fase instantánea del estado inicial.

De acuerdo con la realización preferida, la obtención de la fase α, camino de fases, se realiza por tránsito de fases ponderado en amplitud cuadrática. Por lo que se tiene en cuenta la diferencia de fase de cada pareja de armónicos y la media de las amplitudes.

La virtud de esta estrategia de obtención del valor de la fase α, es el menor error cuadrático entre el difonema real y el sintético que se obtiene, respecto a otras estrategias probadas.

La fase instantánea del tránsito 14, se obtiene aplicado una función de transición del tipo sinusoidal; para la que el error cuadrático medio es el menor entre el resultado sintético de la señal y el patrón real de cuantos se han probado.

A pesar que se ha descrito una realización concreta de la presente invención, es evidente que el experto en la materia podrá introducir variantes y modificaciones, o substituir los detalles por otros equivalentes, sin apartarse del ámbito de protección definido por las reivindicaciones adjuntas.

Por ejemplo, se podrían utilizar otro tipo de camino de fases a seguir para la transición entre las parejas de armónicos de los fonemas. Así como diferentes funciones de tránsito.

REIVINDICACIONES

1. Procedimiento para la síntesis de difonemas y/o polifonemas para la generación de voz sintética **caracterizado** por el hecho de dicha generación se realiza mediante la transición de los fonemas que componen los difonemas y/o polifonemas a generar a partir de un único período de la estructura frecuencial real de los fonemas constituyentes.

2. Procedimiento según la reivindicación 1, **caracterizado** por el hecho de que el tránsito de los fonemas constituyentes se realiza entre estados estacionarios de los fonemas a generar.

3. Procedimiento según la reivindicación 1, **caracterizado** por el hecho de que la transición entre fonemas permite la modificación, si es necesario, de la frecuencia fundamental y de la energía de los fonemas constituyentes.

4. Procedimiento según la reivindicación 3, **caracterizado** por el hecho de que la modificación de la frecuencia fundamental y la energía de los fonemas constituyentes permite, si es necesario, la asignación de características prosódicas a la señal portadora generada.

5. Procedimiento según la reivindicación 1, **caracterizado** por el hecho de que los fonemas sobre los que se transita se caracterizan como coeficientes de Fourier componiendo a cada fonema en una serie de armónicos.

6. Procedimiento según la reivindicación 1 y 5, **caracterizado** por el hecho de que la transición entre fonemas se realiza mediante una transición continua entre parejas de armónicos, pareja compuesta por un armónico del fonema inicial y un armónico del fonema final.

7. Procedimiento según la reivindicación 1, 5 y 6, **caracterizado** por el hecho de que se incorporan armónicos de amplitud cero, si es necesario, al conjunto de armónicos que componen un fonema, si este se encuentra en minoría respecto al otro fonema, hasta completar las parejas de armónicos entre los dos fonemas.

8. Procedimiento según la reivindicación 1 a 7, **caracterizado** por el hecho de que el proceso de la transición entre fonemas se descompone en al menos tres tramos:

(a) primer tramo temporal previo al inicio del tránsito en el que los valores de la señal generada corresponden a los valores del fonema que inicia la transición;

(b) segundo tramo temporal posterior al inicio del tránsito y anterior al final de la transición; y

(c) tercer tramo temporal posterior al final del tránsito en el que los valores de la señal generada corresponden a los valores del fonema que finaliza la transición.

9. Procedimiento según la reivindicación 1 y 8, **caracterizado** por el hecho de que el segundo tramo temporal de la transición entre fonemas, la frecuencia fundamental de la señal generada se encuentre en un valor intermedio de las frecuencias del tramo temporal primero y tercero.

10. Procedimiento según la reivindicación 9, **caracterizado** por el hecho de que en el caso de tener que corregir la frecuencia fundamental de la señal generada para situarla en un valor intermedio de las frecuencias del tramo temporal primero y tercero, la corrección se aplica sumándole un valor de fase a las fases de los armónicos del fonema final.

11. Procedimiento según la reivindicación 9, **caracterizado** por el hecho de que en el caso de tener que corregir la frecuencia fundamental de la señal generada para situarla en un valor intermedio de las frecuencias del tramo temporal primero y tercero, la corrección se aplica sumándole un valor de fase a las fases de los armónicos del fonema inicial.

12. Procedimiento según la reivindicación 9, **caracterizado** por el hecho de que en el caso de tener que corregir la frecuencia fundamental de la señal generada para situarla en un valor intermedio de las frecuencias del tramo temporal primero y tercero, la corrección se aplica restándole un valor de fase a las fases de los armónicos del fonema final.

13. Procedimiento según la reivindicación 9, **caracterizado** por el hecho de que en el caso de tener que corregir la frecuencia fundamental de la señal generada para situarla en un valor intermedio de las frecuencias del tramo temporal primero y tercero, la corrección se aplica restándole un valor de fase a las fases de los armónicos del fonema inicial.

14. Procedimiento según la reivindicación 9, **caracterizado** por el hecho de que en el caso de tener que corregir la frecuencia fundamental de la señal generada para situarla en un valor intermedio de las frecuencias del tramo temporal primero y tercero, la corrección se aplica restándole: un valor de fase a las fases de los armónicos del fonema final y otro valor de a las fases de los armónicos del fonema inicial.

15. Procedimiento según la reivindicación 9, **caracterizado** por el hecho de que en el caso de tener que corregir la frecuencia fundamental de la señal generada para situarla en un valor intermedio de las frecuencias del tramo temporal primero y tercero, la corrección se aplica sumándole: un valor de fase a las fases de los armónicos del fonema final y otro valor de fase a las fases de los armónicos del fonema inicial.

16. Procedimiento según la reivindicación 9, **caracterizado** por el hecho de que en el caso de tener que corregir la frecuencia fundamental de la señal generada para situarla en un valor intermedio de las frecuencias del tramo temporal primero y tercero, la corrección se aplica sumándole un valor de fase a las fases de los armónicos del fonema final y restándole un valor de fase a las fases de los armónicos del fonema inicial.

17. Procedimiento según la reivindicación 9, **caracterizado** por el hecho de que en el caso de tener que corregir la frecuencia fundamental de la señal generada para situarla en un valor intermedio de las frecuencias del tramo temporal primero y tercero, la corrección se aplica restándole un valor de fase a las fases de los armónicos del fonema final y sumándole un valor de fase a las fases de los armónicos del fonema inicial.

18. Procedimiento según la reivindicación 1 a 17, **caracterizado** por el hecho de que la transición entre parejas de armónicos en el tramo temporal segundo, siga al menos, una función de transición de la componente de fase que conforma la señal generada.

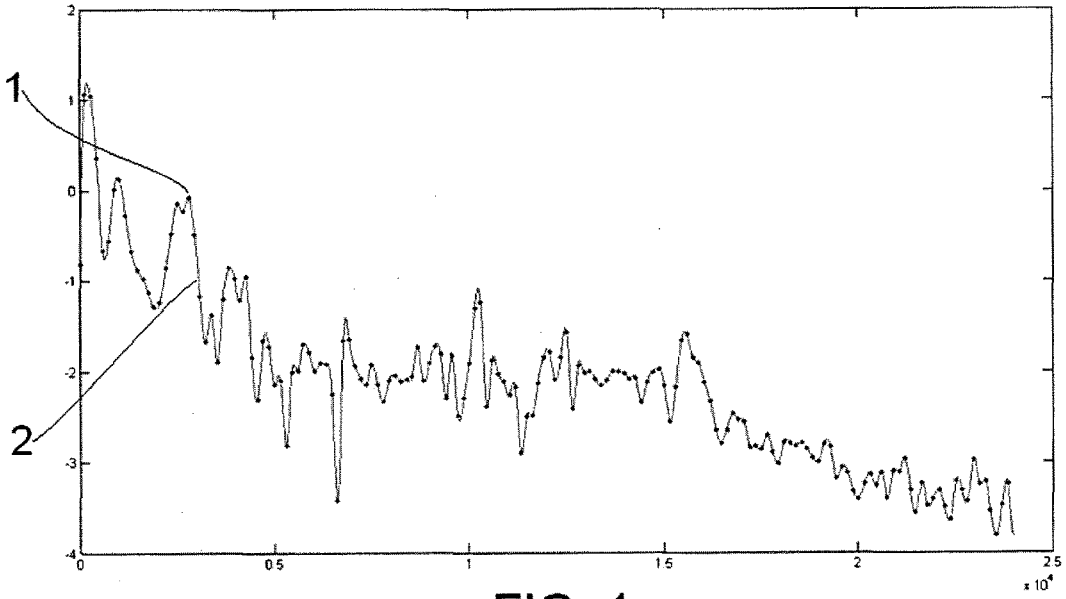
1
2
2
1
0
-1
-2
-3
-4
0
0.5
1
1.5
2
2.5
x 10^4

FIG. 1

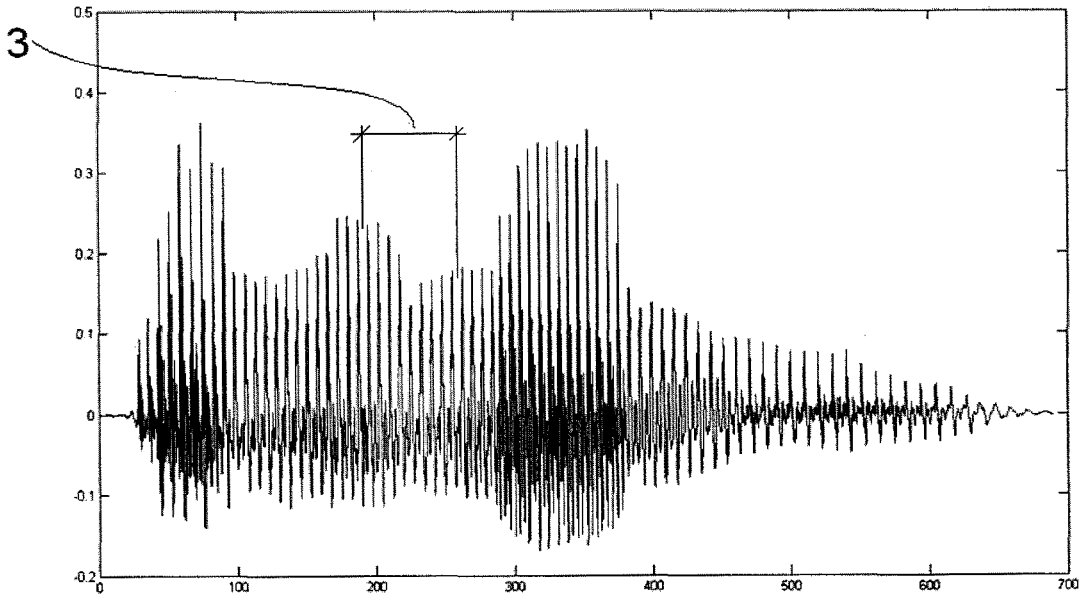
3
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0
100
200
300
400
500
600
700

FIG. 2

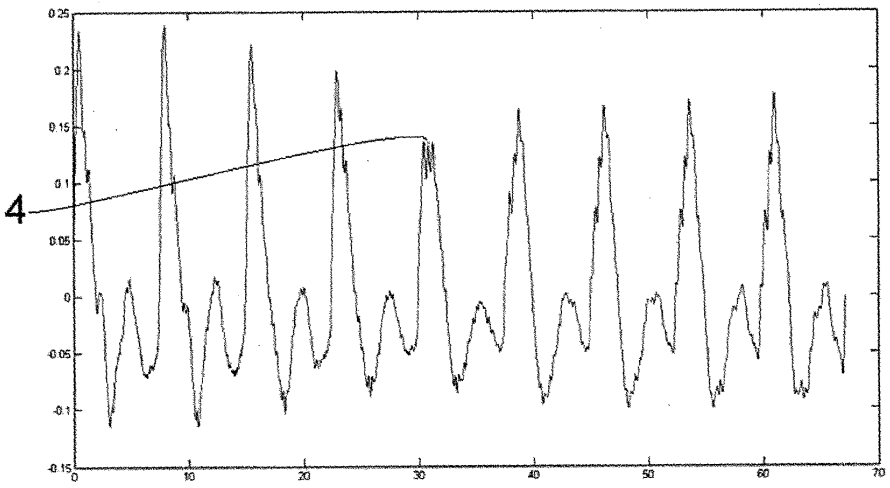

FIG. 3

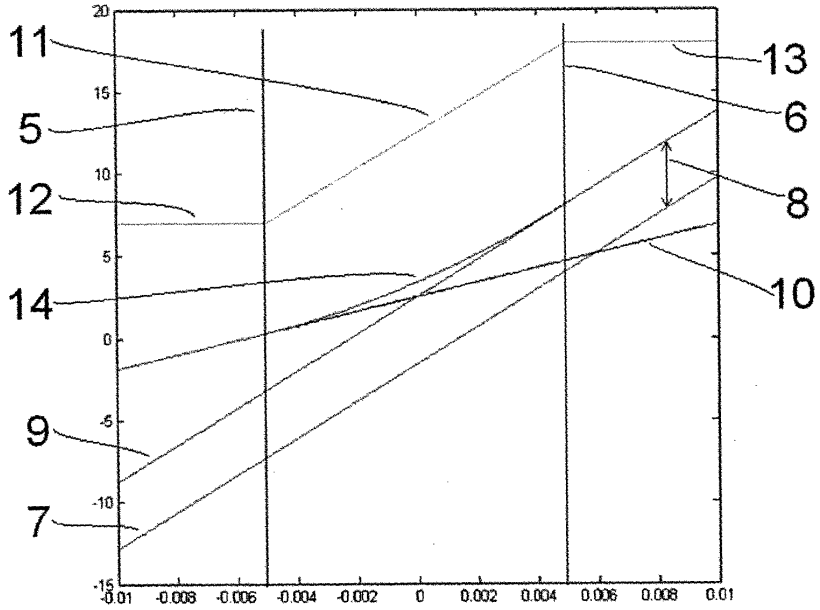

FIG. 4

OFICINA ESPAÑOLA
DE PATENTES Y MARCAS

ESPAÑA

(21) N.º solicitud: 201000256

(22) Fecha de presentación de la solicitud: 23.02.2010

(32) Fecha de prioridad:

INFORME SOBRE EL ESTADO DE LA TECNICA

(51) **Int. Cl. :** Ver Hoja Adicional

DOCUMENTOS RELEVANTES

Categoría	(56) **Documentos citados**	**Reivindicaciones afectadas**
A	WO 9632711 A1 (BRITISH TELECOMM ET AL.) 17/10/1996,	1
A	US 4692941 A (JACKS RICHARD P ET AL.) 08/09/1987,	1
A	US 4601052 A (SAITO HIROSHI ET AL.) 15/07/1986,	1
A	JP 2006084859 A (ATR ADVANCED TELECOMM RES INST) 30/03/2006,	1

Categoría de los documentos citados
X: de particular relevancia
Y: de particular relevancia combinado con otro/s de la misma categoría
A: refleja el estado de la técnica
O: referido a divulgación no escrita
P: publicado entre la fecha de prioridad y la de presentación de la solicitud
E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

El presente informe ha sido realizado

☒ para todas las reivindicaciones ☐ para las reivindicaciones nº:

Fecha de realización del informe	**Examinador**	**Página**
25.05.2012	M. d. González Vasserot	1/4

CLASIFICACIÓN OBJETO DE LA SOLICITUD

G10L13/02 (2006.01)
G10L13/08 (2006.01)
G10L13/06 (2006.01)

Documentación mínima buscada (sistema de clasificación seguido de los símbolos de clasificación)

G10L

Bases de datos electrónicas consultadas durante la búsqueda (nombre de la base de datos y, si es posible, términos de búsqueda utilizados)

INVENES, EPODOC, WPI

Fecha de Realización de la Opinión Escrita: 25.05.2012

Declaración

Novedad (Art. 6.1 LP 11/1986)	Reivindicaciones	1-18	**SI**
	Reivindicaciones		**NO**
Actividad inventiva (Art. 8.1 LP11/1986)	Reivindicaciones	1-18	**SI**
	Reivindicaciones		**NO**

Se considera que la solicitud cumple con el requisito de aplicación industrial. Este requisito fue evaluado durante la fase de examen formal y técnico de la solicitud (Artículo 31.2 Ley 11/1986).

Base de la Opinión.-

La presente opinión se ha realizado sobre la base de la solicitud de patente tal y como se publica.

1. Documentos considerados.-

A continuación se relacionan los documentos pertenecientes al estado de la técnica tomados en consideración para la realización de esta opinión.

Documento	Número Publicación o Identificación	Fecha Publicación
D01	WO 9632711 A1 (BRITISH TELECOMM et al.)	17.10.1996
D02	US 4692941 A (JACKS RICHARD P et al.)	08.09.1987
D03	US 4601052 A (SAITO HIROSHI et al.)	15.07.1986
D04	JP 2006084859 A (ATR ADVANCED TELECOMM RES INST)	30.03.2006

2. Declaración motivada según los artículos 29.6 y 29.7 del Reglamento de ejecución de la Ley 11/1986, de 20 de marzo, de Patentes sobre la novedad y la actividad inventiva; citas y explicaciones en apoyo de esta declaración

Los documentos citados solo muestran el estado general de la técnica, y no se consideran de particular relevancia. Así, la invención reivindicada se considera que cumple los requisitos de novedad, actividad inventiva y aplicación industrial.

1.- El objeto de la presente solicitud de patente se refiere al procedimiento para la generación de la señal acústica de voz sintética de sonidos a partir de una mínima información previa de los fonemas constituyentes obtenidos por segmentación de una grabación previa; dicha información mínima consiste en la envolvente espectral correspondiente a un único periodo situado en la zona más estacionaria del fonema.

2.- El problema planteado por el solicitante es generar la señal acústica de voz sintética sin la necesidad de disponer de una gran cantidad de palabras grabadas previamente con una mejora desde el punto de vista económico y del tiempo de elaboración. El proceso de tránsito entre fonemas deberá preasignar las evolventes de las características prosódicas implícitas en la señal portadora generada, como entonación, intensidad y duración de los fonemas necesario para la transmisión de un mensaje emocional que se acerque en calidad al habla natural. El procedimiento para el tránsito además evoluciona desde la composición frecuencial del fonema de origen al de destino por caminos en los que en ningún momento la percepción sonora es disonante. Para ello durante el tránsito se ha de poder modificar la frecuencia fundamental y la energía de los fonemas constituyentes. También se ha de producir esta transición sin que durante la misma se produzcan discontinuidades audibles.

El documento D1 puede considerarse como el representante del estado de la técnica más cercano ya que en este documento confluyen la mayoría de las características técnicas reivindicadas.

Análisis de la reivindicación 1

D1 se diferencia del documento de solicitud de patente en que en el procedimiento para la síntesis de varios fonemas para la generación de voz sintética no se realiza mediante la transición de los fonemas que componen los polifonemas a generar a partir de un único periodo de la estructura frecuencial real de los fonemas constituyentes.

La reivindicación 1 es nueva (Art. 6.1 LP 11/1986) y tiene actividad inventiva (Art. 8.1 LP11/1986).

Análisis del resto de los documentos

De este modo, ni el documento D1, ni ninguno del resto de los documentos citados en el Informe del Estado de la Técnica, tomados solos o en combinación, revelan la invención en estudio tal y como es definida en las reivindicaciones independientes, de modo que los documentos citados solo muestran el estado general de la técnica, y no se consideran de particular relevancia. Además, en los documentos citados no hay sugerencias que dirijan al experto en la materia a una combinación que pudiera hacer evidente la invención definida por estas reivindicaciones y no se considera obvio para una persona experta en la materia aplicar las características incluidas en los documentos citados y llegar a la invención como se revela en la misma.