(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 364 005**

(51) Int. Cl.:
G10L 21/02 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **08804436 .7**
(96) Fecha de presentación : **19.09.2008**
(97) Número de publicación de la solicitud: **2215632**
(97) Fecha de publicación de la solicitud: **11.08.2010**

(54) Título: **Procedimiento, dispositivo y medio de código de programa informático para la conversión de voz.**

(45) Fecha de publicación de la mención BOPI:
22.08.2011

(45) Fecha de la publicación del folleto de la patente:
22.08.2011

(73) Titular/es: **Asociación Centro de Tecnologías de Interacción Visual y Comunicaciones, VICOMTech**
Mikeletegi Pasealekua, 57
Parque Tecnológico
20009 San Sebastián, ES
María Arantzazu del Pozo Echezarreta

(72) Inventor/es:
Pozo Echezarreta, María Arantzazu del

(74) Agente: **Carpintero López, Mario**

DESCRIPCIÓN

Procedimiento, dispositivo y medio de código de programa informático para la conversión de voz

Campo de la invención

La presente invención se refiere a procedimientos y sistemas para la conversión de voz.

Estado de la técnica

La Conversión de Voz se dirige a la transformación del habla de un hablante fuente para que suene como un hablante diana diferente. Los sintetizadores de texto a voz, sistemas de diálogo y reparación del habla están entre las numerosas aplicaciones que pueden beneficiarse gratamente del desarrollo de la tecnología de la conversión de voz.

Las representaciones de las señales de voz más ampliamente usadas son el Modelo Fuente - Filtro y el Modelo Sinusoidal. La representación Fuente - Filtro (G, Fant, Acoustic Theory of Speech Production, ISBN 9027916004) se basa en un modelo de producción simple compuesto de una onda fuente glotal que excita un filtro variable en el tiempo cargado en su salida por la radiación de los labios. El principal reto en el modelo Fuente - Filtro es la estimación de la onda glotal y los parámetros del filtro del tracto vocal de la señal del habla.

Entre las parametrizaciones existentes de la onda glotal, el modelo Liljencrants-Fant (LF) (The LF-model revisited. Transformation and frequency domain analysis, STL – QPSR, vol. 36, núm. 2 – 3, 1995, pág. 119 – 156) se ha convertido en el modelo de elección para la investigación sobre la fuente glotal. Ha demostrado ser capaz de modelar una amplia gama de fonaciones que se producen de forma natural y los efectos de las variaciones de sus parámetros son fáciles de entender. Explota la linealidad de las propiedades de la invariancia en el tiempo de la representación Fuente - Filtro y asume la conmutación de los filtros del tracto vocal y de la radiación de los labios para combinar el modelo de excitación fuente y la radiación de los labios en la parametrización de la derivada de la onda glotal.

La Predicción Lineal (LP) es una técnica popular usada para obtener una parametrización combinada de los componentes de la fuente glotal, el tracto vocal y de la radiación de los labios en un único filtro omnipolar $H(z)$. Dicho filtro se excita, tal como se muestra en la figura 1, por una secuencia de impulsos separados en el período fundamental To durante el habla sonora y mediante el ruido gaussiano blanco durante el habla no sonora. Si la señal del habla fuera realmente la respuesta de un filtro omnipolar, el error o residual de LP sería un tren de impulsos separados en los instantes de excitación sonora y el modelo de la fuente de voz de impulso/ruido sería exacto. En la práctica, sin embargo, el residuo de LP se parece más a una señal de ruido blanco con valores mayores alrededor de los instantes de excitación. Aunque la excitación del filtro de LP con el residuo de LP da como resultado un habla que es indistinguible del original, usando un tren de impulsos como excitación sonora se produce un habla con una calidad muy zumbante. La fuerza de la LP descansa en su capacidad para estimar automáticamente un conjunto de coeficientes de filtro que representan de forma compacta la cubierta del espectro del habla, haciéndolo popular en aplicaciones en las que las características espectrales de la onda de la voz necesitan ser capturadas con un pequeño número de parámetros. Su principal inconveniente, por otra parte, proviene del modelo sobre–simplificado de la onda glotal que evita su uso en sistemas que requieran salidas de voz de alta calidad.

Como alternativa a la LP, H. Lu y colaboradores han propuesto un procedimiento de optimización convexa para estimar automáticamente el filtro del tracto vocal y la onda glotal conjuntamente (Joint estimation of vocal tract filter and glottal source waveform via convex optimization, Proc, 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, Nueva York, oct. 17 – 20, 1999). El mejor modelo de la fuente glotal empleado por esta aproximación da como resultado un habla que tiene una mejor calidad que la de la LP. Además, la parametrización de la onda glotal permite su modificación paramétrica, que puede explotarse en aplicaciones de conversión de voz.

Los Modelos Sinusoidales asumen que la onda del habla está compuesta de la suma de un pequeño número de sinusoides con amplitudes, frecuencias y fases variables en el tiempo. Dicho modelo fue principalmente desarrollado por McAulay y Quatieri (Speech Analysis/Synthesis Based on a Sinusoidal Representation, IEEE Transactions on Acoustics, Speech and Signal Processing, pág. 744 – 754, 1986) a mediados de los 80 y ha demostrado ser capaz de producir un habla de alta calidad incluso después de transformaciones de tono y escala de tiempo. Sin embargo, a causa del alto número de amplitudes sinusoidales, frecuencias y fases involucradas, el modelo sinusoidal resulta menos flexible que la representación de Fuente - Filtro para modificar características espectrales.

Para obtener un habla convertida de alta calidad, las implementaciones de conversión de voz (VC) de la técnica actual emplean principalmente variaciones y extensiones del modelo sinusoidal original. Además, generalmente adoptan una formulación de Fuente - Filtro basada en la LP para llevar a cabo las transformaciones espectrales.

Las cubiertas espectrales están generalmente codificadas en frecuencias espectrales en línea (LSF) para la conversión de voz, ya que las LSF han demostrado poseer muy buenas características de interpolación lineal y relacionarse bien con la asignación de formatos y ancho de banda. Ya que la resolución de frecuencia del oído humano es mayor a bajas frecuencias que a altas frecuencias, las cubiertas espectrales a menudo se envuelven en una escala no lineal, por

ejemplo, la escala de Bark, tomando en cuenta la sensibilidad no uniforme del oído humano. Habitualmente solo se transforman las cubiertas espectrales de segmentos de habla sonora, ya que los sonidos no sonoros contienen poca información del tracto vocal y sus cubiertas espectrales presentan altas variaciones. Entre las diferentes técnicas de conversión de cubiertas espectrales existentes, se ha hallado que las transformaciones lineales de probabilidad continua son más robustas y eficientes. Estas pueden obtenerse a través de al menos una minimización del error cuadrático de las bases de datos de entrenamiento de fuente y diana o usando marcos generales de transformación de probabilidad máxima (Ye, H. y Young, S. Quality-enhanced Voice Morphing using Maximum Likelihood Transformations, IEEE Audio Speech and Language Processing, vol. 14, núm. 4, pág. 1301 – 1312, 2006). Un problema de la compartición de todos los procedimientos de conversión de cubiertas espectrales es la ampliación de los picos espectrales, la expansión de los anchos de banda formantes y la sobre-suavización provocada por el efecto de promediación de las interpolaciones de los parámetros. El fenómeno hace que el sonido del habla convertida sea ligeramente apagado. Para resolver este asunto, a menudo se aplica un post-filtrado como etapa de post-procesamiento para anchos de banda de formantes estrechos y se suprime el ruido en los valles espectrales como, por ejemplo, en Ye, H. y Young, S. Quality-enhanced Voice Morphing using Maximum Likelihood Transformations, IEEE Audio Speech and Language Processing, vol. 14, núm. 4, pág. 1301 – 1312, 2006.

Como para la conversión de residuos de LP, los sistemas sinusoidales de VC han desarrollado procedimientos de predicción y selección de residuos (D. Suendermann, A. Bonafonte, H. Ney, y J. Hoege, A study on residual prediction techniques for voice conversion, in Proc. ICASSP, 2005, pág. 13 – 16) basados en la correlación entre la cubierta espectral y los residuos de LP. Estos procedimientos reintroducen el detalle espectral diana perdido después de la conversión de la cubierta. Ya que los residuos contienen los errores introducidos por la parametrización de LP, se ha hallado que las técnicas de predicción de residuos mejoran el rendimiento de la conversión. Sin embargo, los residuos de LP no constituyen un modelo exacto de fuente de voz y la predicción de residuos sola no es capaz de modificar la calidad de la fuente de voz. Esto no permite su uso en aplicaciones que requieran modificaciones de la claridad de la voz tales como, por ejemplo, la reparación del habla.

La solicitud de patente WO 2008/018653 A1 presenta una técnica adicional de conversión de voz que usa los parámetros de Liljencrants-Fant de la onda glotal.

Resumen de la invención

Por lo tanto es un objeto de la presente invención suministrar un procedimiento de conversión de voz basado en un modelo de Fuente - Filtro que use una representación de la fuente glotal más exacta que los residuos de LP. Esto permite el uso de transformaciones lineales probabilísticas continuas para la conversión de la fuente de voz.

En particular, es un objeto de la presente invención un procedimiento para convertir una señal de voz de un hablante en una señal de voz convertida, que comprende una etapa de entrenamiento, una etapa de conversión y una etapa de síntesis.

La etapa de entrenamiento comprende, dada una base de datos de entrenamiento de datos fuente y diana paralelos, para cada período de tono de dicha base de datos de entrenamiento: el modelado de cada período de tono por medio de una onda glotal y de un filtro de tracto vocal de acuerdo con el modelo de Lu y Smith, para obtener un conjunto de parámetros de LF, dicho conjunto de parámetros de LF comprende un parámetro de fuerza de excitación y un conjunto de parámetros T que modela una onda glotal, y un conjunto de coeficientes de filtro de tracto vocal omnipolar; que convierten dichos parámetros T en parámetros R; convirtiendo dichos coeficientes de filtro de tracto vocal omnipolar en frecuencias espectrales en línea en la escala de Bark; definiendo un vector glotal a convertir; definiendo un vector de tracto vocal a convertir, comprendiendo dicho vector de tracto vocal dichas frecuencias espectrales en línea en la escala de Bark, aplicando eliminación del ruido de la ondícula para obtener una estimación de un ruido de aspiración glotal.

La etapa de entrenamiento también comprende, a partir de un conjunto de vectores de tracto vocal obtenidos para cada período de tono de dicha base de datos de entrenamiento, la estimación de una función de transformación lineal probabilística continua de tracto vocal que usa el criterio del error cuadrático mínimo.

La etapa previa de modelado comprende además los pasos de modelar dicha estimación de ruido de aspiración modulando el ruido gaussiano de varianza unitaria cercano a cero con la mencionada onda glotal modelada y ajustar su energía para coincidir con la de la mencionada estimación del ruido de aspiración. Además, el vector glotal a convertir comprende dicho parámetro de fuerza de excitación, dichos parámetros R y dicha energía de estimación del ruido de aspiración.

En la etapa de conversión, una onda de voz de prueba se modifica y se transforma en un conjunto de parámetros convertidos.

En la etapa de síntesis, una onda de voz convertida se sintetiza a partir de dicho conjunto de parámetros convertidos.

Preferiblemente, la etapa de entrenamiento comprende además: a partir del conjunto de vectores glotales obtenidos para cada período de tono de la mencionada base de datos de entrenamiento, la estimación de una función de transformación lineal probabilística continua de la onda glotal usando el criterio de error cuadrático mínimo.

El paso de modelado de cada período de tono por medio de una onda glotal y un filtro de tracto vocal de acuerdo con el modelo de Lu y Smith, comprende preferiblemente los pasos de: modelar la onda glotal usando el modelo de Rosenberg-Klatt; usar optimización convexa para obtener un conjunto de parámetros de ondas glotales de Rosenberg-Klatt y todos los coeficientes de filtro de tracto vocal omnipolar, en donde dicho paso de usar optimización convexa comprende un paso de pre-énfasis adaptativo para estimar y eliminar la contribución del filtro de inclinación espectral de la onda del habla antes de la optimización convexa. Además, el paso de modelar cada período de tono por medio de una onda glotal y un filtro de tracto vocal de acuerdo con el modelo de Lu y Smith, comprende además los pasos de: obtener una onda glotal derivada mediante el filtrado inverso de dicho período de tono usando dichos coeficientes de filtro de tracto vocal omnipolar, ajustando dicho conjunto de parámetros de LF a dicha onda glotal derivada filtrada mediante estimación directa y optimización no lineal constreñida.

La etapa de conversión comprende preferiblemente, para cada período de tono de dicha de onda de voz de prueba: la obtención de un vector glotal a convertir, comprendiendo dicho vector glotal un parámetro de fuerza de excitación, un conjunto de parámetros R y la energía de dicha estimación de ruido de aspiración; la obtención el vector de tracto vocal a convertir, comprendiendo dicho vector de tracto vocal un conjunto de frecuencias espectrales en línea en la escala de Bark; la aplicación de dicha función de transformación lineal probabilística continua de tracto vocal estimada durante la etapa de entrenamiento para obtener un vector de parámetros de tracto vocal convertidos; la transformación de dicho vector glotal usando dicha función de transformación lineal probabilística continua de la onda glotal estimada durante la etapa de entrenamiento obteniendo así un vector glotal convertido que comprende un conjunto de parámetros convertidos.

En particular, esas etapas de obtención de un vector glotal a convertir y de un vector de tracto vocal a convertir comprenden además los pasos de: modelar cada período de tono por medio de una onda glotal y un filtro de tracto vocal de acuerdo con el modelo de Lu y Smith, para obtener un conjunto de parámetros de LF, dicho conjunto de parámetros de LF comprende un parámetro de fuerza de excitación y un conjunto de parámetros T que modelan una onda glotal y un conjunto de coeficientes de filtro de tracto vocal omnipolar; convertir dichos coeficientes de filtro de tracto vocal omnipolar en frecuencias espectrales en línea en la escala de Bark; convertir dichos parámetros T en parámetros R; definir un vector glotal a convertir; y definir un vector de tracto vocal a convertir.

Preferiblemente, la etapa de conversión comprende además un paso de post-filtrado de dicho vector de parámetros de tracto vocal convertido.

La etapa de síntesis, en la cual dicha onda de voz convertida se sintetiza a partir de dicho conjunto de parámetros convertidos, comprende preferiblemente los pasos de: interpolar las trayectorias de dichos parámetros convertidos de cada período de tono, obteniendo así un conjunto de parámetros interpolados que comprenden parámetros R interpolados, energía interpolada y un vector de tracto vocal interpolado, convertir dicho vector de tracto vocal interpolado en un vector de coeficientes de filtro omnipolar; convertir dichos parámetros R interpolados en parámetros T interpolados; para cada marco de dicha onda de voz de prueba, generando una señal de excitación.

Preferiblemente, la etapa de generar una señal de excitación comprende, para cada uno de dichos marcos: si dicho marco es sonoro: a partir de dichos parámetros T interpolados y de dicho parámetro de fuerza de excitación, generar una onda glotal; a partir de dicho parámetro de energía de ruido de aspiración interpolado generar ruido de aspiración interpolado; generar dicha señal de excitación sonora añadiendo dicha onda glotal interpolada y dicho ruido de aspiración interpolado. Y, si dicho marco no es sonoro: generar dicha señal de excitación no sonora a partir de una fuente de ruido gaussiano.

Además, la etapa de síntesis comprende adicionalmente: la generación de una contribución sintética de cada marco filtrando dicha señal de excitación con dicho vector de coeficientes de filtro omnipolar, la multiplicación de dicha contribución sintética mediante una ventana de Hamming, la superposición y la adición, para generar la señal de voz convertida.

La presente invención también suministra un procedimiento aplicable a transformaciones de calidad de voz, tales como la reparación del habla traqueo-esofágica, que comprende al menos algunos de los pasos del procedimiento antes mencionado.

Otro objeto de la presente invención es suministrar un dispositivo que comprende medios para llevar a cabo el procedimiento antes mencionado.

Finalmente, un objeto adicional de la presente invención es suministrar medios de código de programa informático adaptados para realizar los pasos del procedimiento previamente mencionado cuando dicho programa se ejecuta en un ordenador, un procesador de señales digitales, una matriz de puertas programables por campo, un circuito integrado específico para la aplicación, un microprocesador, un microcontrolador o cualquier otra forma de hardware programable.

Las ventajas de la invención propuesta serán evidentes en la descripción siguiente.

Breve descripción de los dibujos

Para completar la descripción y para proporcionar una mejor comprensión de la invención, se suministra un conjunto de dibujos. Dichos dibujos forman 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, al contrario, como ejemplo de cómo puede realizarse la invención. Los dibujos comprenden las siguientes figuras:

La figura 1 muestra un diagrama esquemático convencional del modelo de LP.

La figura 2 muestra un diagrama esquemático del modelo de síntesis de análisis de estimación conjunta (JEAS) de acuerdo con una realización de la presente invención.

La figura 3 muestra un diagrama esquemático del modelado de la onda glotal.

La figura 4 muestra un diagrama esquemático del modelado de la onda glotal derivada.

La figura 5 muestra impulsos de LF típicos que se corresponden con las ondas glotal y glotal derivada.

La figura 6 muestra un modelo convencional de la fuente de voz.

La figura 7 muestra un ejemplo de estimación conjunta: a) período de habla, b) espectro de habla y cubierta espectral conjuntamente estimada, c) residuo filtrado inverso y onda RK conjuntamente estimados.

La figura 8 muestra una onda glotal derivada RK.

La figura 9 muestra un diagrama esquemático de un modelo convencional de la inclinación espectral.

La figura 10 muestra los efectos del pre-énfasis adaptativo.

La figura 11 muestra un ejemplo de ajuste de LF en fonaciones normal, entrecortada y forzada.

La figura 12 muestra un resultado típico de eliminación de ruido.

La figura 13 muestra parámetros del modelo del ruido de aspiración estándar.

La figura 14 muestra la modulación del ruido gaussiano mediante una onda LF.

La figura 15 muestra un diagrama esquemático de una aproximación de modelado del ruido de aspiración de acuerdo con una realización de la presente invención.

La figura 16 muestra un diagrama esquemático del esquema de síntesis de superposición – adición empleado.

La figura 17 ilustra el nuevo muestreo del contorno del tamaño del marco.

La figura 18 muestra las cubiertas espectrales de JEAS frente a PSHM.

La figura 19 muestra la transformación lineal probabilística continua de las ondas glotales LF

La figura 20 muestra las relaciones de distorsión R_{LSF} de las cubiertas espectrales de PSHM y de JEAS convertidas.

La figura 21 muestra las relaciones de distorsión de R_{LSD} de los espectros predicho residual (RP) y convertido de onda glotal (GWC).

La figura 22 muestra los resultados de la prueba ABX.

La figura 23 muestra el resultado de la prueba de comparación de calidad.

Descripción detallada de la invención

Definiciones

En el contexto de la presente invención, el término “aproximadamente” y los términos del subgrupo (tales como “aproximado”, “aproximación”, etc) deben entenderse como valores o formas indicativas muy cerca de aquellos que acompañan al término antes mencionado. Es decir, puede aceptarse una desviación de un valor exacto dentro de límites razonables, ya que cualquier experto en la técnica entenderá que dicha desviación de los valores o formas indicados es inevitable debido a las inexactitudes de las mediciones, etc. Lo mismo se aplica al término “cercano”.

En el contexto de la presente invención, los siguientes términos se definen como sigue:

La expresión “periodo del tono” significa un segmento de una onda de voz que comprende un período de la frecuencia fundamental.

El término “marco” significa un segmento de una onda de voz, que se corresponde con el período del tono en las partes sonoras y con una cantidad fija de tiempo en las partes no sonoras. En una realización preferida de la presente invención, que no debe interpretarse como una limitación de la presente invención, un marco se corresponde con 10 milisegundos en las partes no sonoras.

La expresión “datos fuente” se refiere a una colección de ondas de voz lanzadas por un hablante fuente, mientras la expresión “datos diana” se refiere a una colección de ondas de voz lanzadas por un hablante diana. Además, la expresión “datos fuente y diana paralelos” se refiere a una colección de ondas de voz lanzadas por los hablantes tanto fuente como diana.

En este texto, el término “comprende” y sus derivados (tales como “comprendiendo”, etc) no debe entenderse en un sentido excluyente, es decir, estos términos no deben interpretarse como que excluyen la posibilidad de que lo que se describe y define pueda incluir elementos o pasos adicionales.

1. Síntesis del Análisis de estimación conjunta

A continuación se describe un procedimiento de modelado del habla para el análisis, modificación y síntesis del habla. El modelo se denomina síntesis de análisis de estimación conjunta (JEAS). Su mayor ventaja es la parametrización automática y simultánea del tracto vocal y de la fuente de voz, que permite la manipulación no solo de las cubiertas espectrales, sino también de características glotales. Además, también soporta tono de alta calidad y modificaciones en la escala de tiempo. A continuación, se describe el modelo de fuente de voz empleado y la técnica de deconvolución del filtro de la fuente y se implementan el análisis de modo, transformaciones de síntesis y prosódicas.

1.1 Modelo del habla

La figura 2 muestra un diagrama esquemático del modelo JEAS. Se basa en una representación general de Fuente - Filtro. Emplea ruido gaussiano blanco y gaussiano blanco modulado en amplitud para modular los componentes del *ruido de turbulencia y aspiración* respectivamente, Un diferenciador para la *Radiación de los Labios* y un Filtro omnipolar para representar *el Tracto Vocal*. Además, se adopta el modelo Liljencrants – Fant (LF) para capturar mejor las características de la onda glotal derivada. Entonces, para estimar las diferentes parametrizaciones de los componentes del modelo de la onda del habla, se aplica una técnica de estimación de los parámetros de la fuente de voz de unión y del tracto vocal basándose en la Optimización Convexa.

Después, se explica el modelado de la fuente de voz.

Se han propuesto numerosos modelos paramétricos de la fuente glotal en la literatura. A pesar de sus diferencias, todos comparten muchas características comunes y pueden describirse mediante un pequeño conjunto de parámetros. En la mayoría de los casos, explotan las propiedades de linealidad e invarianza del tiempo en la representación de Fuente - Filtro y asumen la conmutación de los filtros del tracto vocal y de la radiación de los labios para combinar el modelo de la excitación fuente y de la radiación de los labios en la parametrización de la *derivada de la onda glotal* tal como se muestra en la figura 4.

El presente procedimiento adopta el modelo LF bien conocido, que es un modelo de dominio de tiempo de cuatro parámetros de un ciclo de la onda glotal derivada. Los impulsos típicos LF que se corresponden con las ondas glotal y glotal derivada se muestran en la figura 5. Matemáticamente, puede describirse como

$$g(n) = \begin{cases} E_0 e^{\alpha n} \sin(\omega_g n) & 0 \le n < T_e \\ -\frac{E_e}{\varepsilon T_a}\left[e^{-\varepsilon(n-T_e)} - e^{\varepsilon(T_c-T_e)}\right] & T_e \le n < T_c \end{cases} \quad (1)$$

El modelo consta de dos segmentos: el primero caracteriza la onda glotal derivada a partir del instante de la apertura glotal hasta el instante de excitación principal T_e, donde la amplitud alcanza el máximo valor negativo $-E_e$.

Según se muestra en la ecuación 1, el segmento es una función sinusoidal que crece exponencialmente en amplitud,

$$F_g = \frac{\omega_g}{2\pi}$$

Siendo la frecuencia de la función del seno y determinando α la tasa de incremento de amplitud. E_0 es un factor de escala usado para asegurar que la señal esté próxima a cero. El parámetro de sincronización T_p se relaciona con la frecuencia sinusoidal a través de

$$T_P = \frac{1}{2F_g}$$

y denota el instante de máximo flujo glotal. E_e está estrechamente relacionado con la fuerza de la excitación de la fuente y la determinante principal de la intensidad de la señal del habla. Su variación afecta las amplitudes armónicas totales, excepto a los componentes muy bajos que están más determinados por la forma del impulso.

El segundo segmento modela la fase de cierre o retorno desde la excitación principal T_e hasta el instante de cierre total T_c usando una función exponencial. La duración de la fase de retorno se determina de esta forma mediante T_c-T_e. El parámetro principal que caracteriza este segmento es T_a, que representa la "duración efectiva" de la fase de retorno. Esta se define mediante la duración desde T_e hasta el punto en el que una tangente ajustada en el inicio de la fase de retorno cruza cero. ε^{-1} es la constante de tiempo de la función exponencial y puede determinarse iterativamente a partir de T_a, T_e y T_c a través de

$$\varepsilon = \frac{1}{T_a}\left(1 - e^{-\varepsilon(T_c - T_e)}\right).$$

T_o se corresponde con el período fundamental. Generalmente T_c se hace coincidir con la apertura del siguiente impulso. Este hecho podría sugerir que el modelo no toma en cuenta la fase cerrada de la onda glotal. Sin embargo, para valores razonablemente pequeños de T_a la función exponencial ajustará muy cerca de la línea del cero que suministra una fase cerrada sin la necesidad de parámetros de control adicionales.

Junto con la fuerza de excitación E_e, el impulso LF puede determinarse únicamente mediante los parámetros T (T_p, T_e, T_a, T_c). Estos parámetros pueden identificarse fácilmente a partir de la onda glotal derivada estimada. Por lo tanto, se obtienen generalmente primero y entonces los parámetros de síntesis, a partir de los cuales puede calcularse la onda LF directamente, (E_0, α; ω_g; ε) se derivan tomando en cuenta las siguientes restricciones:

$$\int_T^0 g(t) \cdot dt = 0 \qquad (2)$$

$$\omega_g = \frac{\pi}{T_p} \qquad (3)$$

$$\varepsilon T_a = 1 - e^{-\varepsilon(T_c - T_e)} \qquad (4)$$

$$E_0 = -\frac{E_e}{e^{\alpha T_e} \sin(\omega_g T_e)} \qquad (5)$$

Otro conjunto importante de parámetros LF son los parámetros R (R_g, R_k, R_a), que se normalizan respecto a T_0 y se correlacionan con los fenómenos glotales más sobresalientes, es decir, la anchura del impulso glotal y la asimetría y brusquedad del cierre

$$R_g = \frac{T_0}{2 \cdot T_p}; R_k = \frac{T_e - T_p}{T_p}; R_a = \frac{T_a}{T_0}$$

(6)

R_g es una versión normalizada de la frecuencia del formante glotal F_g que se define como la inversa de dos veces la duración de la fase de apertura T_p. R_k es el parámetro de LF que captura la asimetría glotal. Se define como la razón entre las veces de apertura y cierre de las ramas del impulso glotal, y cuanto más grande sea su valor, más simétrico es el impulso. La relación entre R_g, R_k y el Cociente de Apertura OQ es: OQ = (1 + R_k) / (2 R_g). Así, OQ se correlaciona positivamente con R_k y se correlaciona negativamente con R_g. El parámetro R_a se corresponde con el "tiempo de retorno" efectivo T_a normalizado por el período fundamental y captura las diferencias relativas a la inclinación espectral.

Después, se aplica el procedimiento adoptado para la deconvolución de la fuente glotal y del filtro de tracto vocal:

El objetivo de la deconvolución de Fuente - Filtro es obtener estimaciones de los componentes de la fuente glotal y del filtro del tracto vocal a partir de la onda del habla. Existen dos aproximaciones principales de deconvolución. Antes de que se desarrollaran los modelos paramétricos de la onda glotal, el Filtrado Inverso (IF) era el procedimiento de deconvolución más comúnmente empleado. Se basa en el cálculo de una función de transferencia del filtro de tracto vocal, cuya inversa se utiliza para obtener la estimación de la onda glotal que luego puede parametrizarse.

Un enfoque diferente comprende el modelado tanto de la fuente glotal como del filtro del tracto vocal, y de técnicas de desarrollo para estimar conjuntamente los parámetros de los modelos de la fuente y del tracto a partir de la fuente del habla. Los procedimientos de Estimación conjunta son completamente automáticos. Esta es una condición importante que un modelo matemático orientado al análisis, síntesis y modificación de la señal del habla debe satisfacer. Debido a las características de la fuente de voz matemática y de las descripciones del tracto vocal, dicha aproximación es un problema complejo no lineal. Por esta razón, el uso de la LP se ha desarrollado más ampliamente que un método más simple para obtener una parametrización de Fuente - Filtro directa y eficiente de la señal de voz. Su pobre modelado de la fuente de voz no ha limitado su aplicación en el contexto de la codificación del habla y representa eficientemente el espectro del habla con un pequeño número de parámetros. Sin embargo, se ha evitado su uso en aplicaciones de síntesis y transformación del habla. Los avances en la conversión de voz y en la síntesis del habla del Modelo Hidden Marco (HMM) en los últimos años ha enfatizado la importancia de la codificación redefinida de la voz, y así, ha ganado un renovado interés el problema de la estimación conjunta automática de los parámetros de fuente de voz y filtro de tracto vocal.

El procedimiento empleado para obtener los parámetros de fuente de voz JEAS y del modelo del tracto vocal a partir de la onda del habla sigue la segunda aproximación de deconvolución y se basa en la estimación conjunta del filtro del tracto vocal y de la onda glotal propuesta por Lu y Smith (Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, Nueva York, Oct. 17 – 20, 1999).

1.2 Análisis dentro del Modelo de Síntesis de Análisis de Estimación Conjunta

Durante el análisis, los segmentos del habla sonoros y no sonoros se procesan diferentemente debido a sus diversas características de fuente. Mientras que la fuente sonora en el habla sonora se representa mediante una combinación de LF y de modelos del ruido de aspiración, el ruido gaussiano blanco se utiliza para excitar el filtro de tracto vocal en los marcos no sonoros (consulte la figura 2). Su diferente modelado requiere un paso de pre-procesamiento en el que se determinan las secciones de voz sonoras y no sonoras y los instantes de cierre glotal (GCI) de los segmentos de voz. Entonces, se obtienen los parámetros de fuente de voz y de tracto vocal a través de una estimación de Fuente - Filtro conjunta y una reparametrización LF en las secciones sonoras (V) y a través de la LP de autocorrelación estándar y la energía del ruido gaussiano que concuerdan en las partes no sonoras (U).

Se usa un algoritmo, tal como el bien conocido algoritmo de pendiente de fase proyectada de programación dinámica (DYPSA) para la estimación de GCI. Se emplea la función de retardo de grupo en combinación con un procedimiento de proyección de pendiente de fase para determinar los candidatos GCI, más programación dinámica N-best para seleccionar los candidatos más probables de acuerdo con una función de coste que toma en cuenta la similitud de la onda, la desviación del tono, la energía normalizada y la desviación de la pendiente de fase ideal.

La decisión de reproducción se toma basándose en la información de la energía, el cruce por cero y GCI. Los segmentos sonoros se procesan entonces en sincronía con el tono mientras que los marcos no sonoros se extraen periódicamente. En una realización particular, se extraen cada 10 milisegundos.

El procedimiento empleado por la invención para obtener los parámetros de los modelos de la fuente de voz JEAS y del

tracto vocal comprende el uso de un modelo de fuente de voz lo suficientemente simple como para permitir que la deconvolución del filtro de la fuente se formule como un problema de Optimización Convexa. Entonces, la onda glotal derivada obtenida mediante filtrado inverso (IF) con los coeficientes de filtro estimados se reparametriza mediante ajuste del modelo LF.

El éxito de la técnica descansa en el suministro de una restricción de la onda glotal derivada cuando se estima el filtro de tracto vocal. A causa de ello, la onda glotal derivada del IF resultante está más cerca de la excitación glotal verdadera y su ajuste con un modelo LF es menos susceptible de errores.

El algoritmo de estimación conjunta modela la fuente de voz usando el modelo bien conocido de Rosenberg – Klatt (RK), que consiste en una onda de reproducción básica que describe la forma de la onda glotal derivada y un filtro de paso bajo, $\frac{1}{1-\mu^{-1}}$, con μ > 0, según muestra en la figura 6. La derivada RK de la onda glotal viene dada por

$$\hat{g}(n) = \begin{cases} 0 & 1 \le n < n_c \\ 2a(n-n_c) - 3b(n-n_c)^2 & n_c \le n < T_0 \end{cases} ,$$

(7)

donde To se corresponde con el período de tono y n_c se representa la duración de la fase cerrada, que también puede expresarse como

$$n_c = T_0 - \mathrm{OQ} \cdot T_0 ,$$

(8)

siendo OQ el cociente abierto, es decir, la fracción del período de tono en el cual está abierta la glotis. Además, los parámetros *a* y *b* necesitan ser siempre positivos y mantener la siguiente relación,

$$a = b \cdot \mathrm{OQ} \cdot T_0 ,$$

(9)

para mantener una forma de onda adecuada.

La deconvolución de Fuente - Filtro a través de la optimización convexa se realiza minimizando el error cuadrático entre las ondas glotales moderadas y derivadas verdaderas. La onda glotal derivada modelada ĝ(n) se corresponde con la de la ecuación 7, mientras que la onda glotal derivada verdadera g(n) se obtiene a través de un filtrado inverso tal como

$$g(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) ,$$

(10)

donde s(n) es la onda del habla y α_k son los coeficientes del filtro omnipolar del tracto vocal.

El error entre las ondas glotales modelada y derivada verdadera e(n) puede calcularse sustrayendo las ecuaciones (7) y (10)

$$\begin{aligned} e(n) &= \hat{g}(n) - g(n) \\ &= \begin{cases} 0 - s(n) + \sum_{k=1}^{p} \alpha_k s(n-k) & 1 \le n < n_c \\ 2a(n-n_c) - 3b(n-n_c)^2 - s(n) + \sum_{k=1}^{p} \alpha_k s(n-k) & n_c \le n < T_0 \end{cases} \end{aligned}$$

(11)

Reordenando la

expresión anterior y reescribiéndola en forma de matriz tenemos

$$E = \begin{bmatrix} e(1) \\ \vdots \\ e(n_c) \\ e(n_c+1) \\ \vdots \\ e(T_0) \end{bmatrix} = \begin{bmatrix} s(0) & \cdots & s(-p) & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s(n_c-1) & \cdots & s(n_c-p) & 0 & 0 \\ s(n_c) & \cdots & s(n_c+1-p) & 2(1) & -3(1)^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s(T_0-1) & \cdots & s(T_0-p) & 2(T_0-n_c) & -3(T_0-n_c)^2 \end{bmatrix} X - \begin{bmatrix} s(1) \\ \vdots \\ s(n_c) \\ s(n_c+1) \\ \vdots \\ s(T_0) \end{bmatrix}$$

$$= \begin{bmatrix} f_1^T \\ \vdots \\ f_{T_0}^T \end{bmatrix} X - \begin{bmatrix} s_1 \\ \vdots \\ s_{T_0} \end{bmatrix} = FX - S \ ,$$

donde $X = [\alpha_1 \ldots \alpha_p \text{ a b}]^T$ es el vector de parámetros a estimar de manera que se minimice la suma de los cuadrados del error E de la ecuación, es decir

$$\min_X \| E \|^2 = \min_X \sum_{n=1}^{T_0-1} (E(n))^2 = \min_X \| FX - S \|^2 \quad .$$

(12)

H. Lu y colaboradores demostraron (Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, Nueva York, Oct. 17 – 20, 1999) que la simplicidad del modelo glotal RK garantiza que esta optimización sea convexa, es decir, que tenga solamente un mínimo que se corresponde con la solución óptima, y así, pueda resolverse de forma eficiente a través de Programación Cuadrática. Un problema cuadrático se define como sigue

$$\min_X q(X) = \frac{1}{2} X^T H X + g^T X$$

$$\text{subject to : } \quad AX \geq b$$
$$A_{eq} X = b_{eq}$$

(13)

La ecuación (12) puede resolverse usando programación cuadrática si se expande para que tenga su forma, es decir

$$\begin{aligned} \min_x \| FX - S \|^2 &= (FX - S)^T (FX - S) \\ &= X^T F^T F X - 2 S^T F X + S^T S \ , \end{aligned}$$

(14)

definiendo

$$H = 2F^T F$$
$$g^T = -2S^T F$$

(15)

e ignorando el término S^TS, que siempre es positivo, para propósitos de minimización. Además, la ecuación 9 impone las siguientes restricciones de igualdad y desigualdad

$$a > 0$$
$$b > 0$$
$$a = b \cdot \mathrm{OQ} \cdot T_0 \ .$$

(16)

El programa cuadrático derivado puede resolverse usando un número de algoritmos numéricos iterativos existentes. En la implementación desarrollada se ha empleado la función de programación cuadrática del MATLAB Optimization Toolbox. El resultado del problema de minimización es la estimación simultánea de los parámetros del modelo RK *a* y *b* y los coeficientes de filtro omnipolar α_k. La figura 7 muestra un ejemplo de estimación conjunta para un período de tono.

El proceso de estimación conjunta descrito asume que las fases cerrada y abierta están definidas, aunque en la práctica el parámetro que delimita el final de la fase cerrada y el inicio de la fase abierta n_c se desconoce. Su valor óptimo se halla muestreando uniformemente los posibles valores de n_c (empíricamente mostrados para que varíen entre un 0% y un 60% del período del tono T_o), resolviendo el problema cuadrático en cada valor de n_c muestreado y seleccionando la estimación que da como resultado el error mínimo.

Como puede verse en la figura 8, la onda básica de reproducción RK de la ecuación (7) no modela explícitamente la fase de retorno de la onda glotal derivada y cambia bruscamente en los instantes del cierre glotal. Por esta razón, se añade un filtro de paso bajo al modelo básico, con el propósito de reducir la brusquedad del cierre glotal. En el dominio de frecuencias, el coeficiente de filtro μ es responsable del control de la inclinación del espectro de la fuente.

Para permitir la formulación del problema de optimización convexa, el filtro de inclinación espectral se separa del modelo de la fuente y se incorpora en el modelo del tracto vocal añadiendo un polo extra al filtro omnipolar tal como se muestra en la figura 9. Esto implica que los coeficientes del filtro de tracto vocal estimados usando esta fórmula también codifican la información de la pendiente espectral de la fuente de voz. Como resultado, las ondas glotales derivadas obtenidas usando esta aproximación fallan en la captura adecuada de las variaciones de la fase de retorno de la fuente glotal.

La presente invención utiliza pre-énfasis adaptativo para estimar y eliminar la contribución del filtro de inclinación espectral de la onda del habla antes de la optimización convexa. Para que se aplique un análisis de LP e IF para estimar y eliminar la pendiente espectral de los marcos de voz bajo análisis, el efecto del pre-énfasis adaptativo se ilustra en la figura 10: a) espectro del habla y cubierta espectral estimada, b) onda glotal derivada IF y onda LF ajustada, c) espectro de onda glotal derivada y espectro de onda LF ajustada. Las estimaciones de la cubierta del filtro del tracto vocal obtenidas de esta forma no codifican las características de inclinación espectral de la fuente, que se reflejan en la fase de cierre de las ondas glotales derivadas resultantes. Esto mejora el ajuste de la fase de retorno del modelo LF y así, de las altas frecuencias de la fuente glotal.

El modelo LF es capaz de describir más exactamente la onda de la derivada glotal que el modelo RK. Sin embargo, su formulación no lineal más compleja no cumple la condición de convexidad y evita su uso en el algoritmo de estimación de parámetros conjuntos de la fuente de voz y del filtro del tracto vocal. Al contrario, el modelo RK se emplea durante la deconvolución de la Fuente - Filtro y el modelo LF se usa entonces para reparametrizar la onda glotal derivada obtenida mediante filtrado inverso de la onda del habla con los coeficientes de filtro conjuntamente estimados.

El ajuste del modelo LF se lleva a cabo en dos pasos: primero, las estimaciones iniciales de los parámetros T de LF (T_p, T_e, T_a, T_c) y la fuerza de excitación glotal E_e se obtienen a partir de la onda de la fuente de voz IF del dominio de tiempo mediante procedimientos de estimación directos convencionales. Entonces, se redefinen sus valores usando la técnica de optimización no lineal constreñida convencional. El procedimiento completo es como sigue.

Primero se localizan la fuerza de excitación glotal E_e y su índice de tiempo T_e hallando el mínimo de la onda glotal derivada IF. Entonces, se determinan T_p y T_c como los primeros cruces por cero antes y después de T_e respectivamente. Se estima T_a como $T_a = (T_c - T_e)$ 2/3. Se redefinen T_p y T_a usando minimización no lineal constreñida. Ya que las estimaciones iniciales de E_e, T_e y T_c son bastantes fiables, sus valores se mantienen sin cambios durante su optimización. T_a se restringe para que varíe entre 0 y $T_c - T_e$ y T_p hasta b dentro de ± 20% de su estimación inicial. Las fases de retorno y apertura se optimizan separada y secuencialmente. En ambos casos, la función de minimización es la suma del error cuadrático entre la onda glotal derivada IF y la estimación ajustada para la fase particular. La figura 11 muestra un ejemplo de ajuste de LF en fonaciones normal, entrecortada y forzada.

Ya que la parametrización de LF no modela el ruido de aspiración glotal, el componente estocástico presente en la onda glotal derivada IF no se captura durante el ajuste de IF. Sin embargo, perceptualmente, la falta de ruido de aspiración da como resultado una calidad del habla no natural y así, se ha desarrollado una metodología para su extracción y parametrización dentro del marco JEAS.

Se usa la eliminación del ruido de la ondícula para extraer el ruido de aspiración glotal de la estimación de la onda glotal

derivada IF. En una realización preferida, la técnica de eliminación del ruido de la ondícula que se utiliza es el Análisis de Paquetes de Ondícula, que se ha hallado que obtiene estimaciones de ruido de aspiración más fiables en comparación con otras técnicas empleadas para identificar y separar los componentes periódicos y aperiódicos de las señales cuasi-periódicas, tales como el análisis de transformación de frecuencia o la predicción periódica. El Análisis de Paquetes de Ondícula se realiza preferiblemente a nivel 4 con la ondícula de Daubechies de 7º orden, usando umbralización virtual y el criterio de evaluación de umbral de la Estimación de Riesgos Insesgada de Stein. La figura 12 muestra un resultado típico de eliminación de ruido: a) onda glotal derivada IF original y con el ruido eliminado, b) estimación de ruido.

Una vez que se ha extraído una estimación del ruido de aspiración, es necesario que se parametrice. Estudios sobre el ruido de aspiración han mostrado que éste es síncrono con la onda glotal y es probable que presente ráfagas de ruido en el cierre glotal y a menudo también en la apertura glotal. La mayoría de los modelos descuida la naturaleza del impulso de apertura glotal y se aproximan a él como ruido gaussiano modulado en amplitud síncrono con el tono, con mayor energía cerca de los instantes de cierre glotal. La amplitud de la ráfaga de ruido se modula habitualmente usando ventanas Rectangulares, de Hanning o de Hamming. Algunas veces se incluye un filtro de conformación espectral para tener en cuenta la densidad espectral media del ruido de aspiración y el filtrado de paso alto introducido por la conmutación de los filtros de tracto vocal y de radiación. Sin embargo, algunos modelos también desatienden el filtro de conformación espectral ya que se ha encontrado que no es perceptualmente importante. Estas aproximaciones de ruido gaussiano modulado en amplitud síncrono en el tono requiere la determinación de los siguientes parámetros ilustrados en la figura 13 del componente del ruido de aspiración:

-*Umbral mínimo del ruido (N_f)*: el umbral mínimo del ruido de aspiración;

-*Amplitud del impulso del ruido (NP_a)*: el índice de modulación de amplitud del impulso del ruido

-*Posición del impulso del ruido (NP_p)*: la posición del centro de la ventana del impulso del ruido en el período glotal

-*Anchura del impulso del ruido (NP_w)*: la anchura de la ventana del impulso del ruido.

Desafortunadamente, el cálculo automático de los anteriores parámetros a partir de los componentes del ruido de aspiración estimados es problemático en muchos casos. Para evitar estos errores, en la presente invención se sigue una aproximación diferente y, en particular, en la implementación JEAS. Aunque todavía la aproximación al componente de aspiración es el ruido gaussiano modulado en amplitud síncrono con el tono, se emplea una función alternativa que no requiere la estimación de N_f, Np_a, NP_p y NP_w para modular su amplitud: la onda LF. De hecho, la forma de la onda LF sigue la mayoría de las características de modulación de amplitud saliente del ruido de aspiración glotal, es decir, la magnitud de su amplitud aumenta durante la fase abierta y es máxima en el cierre glotal. Si se modula ruido gaussiano estacionario con una onda LF, la señal resultante presentará las dos ráfagas probables de ruido de aspiración cerca de la apertura glotal y del cierre glotal según se muestra en la figura 14 (a) fuente de ruido gaussiano, b) onda LF, c) ruido gaussiano modulado). De acuerdo con pruebas de audición informal, esta aproximación es comparable a las previamente descritas técnicas de modelado basadas en ventanas.

Así, la presente invención, la estimación del ruido de aspiración obtenida para un período de tono particular durante el análisis JEAS se parametriza como sigue. Primero, se modula el ruido gaussiano de varianza unitaria próximo a cero con la onda LF ya ajustada para ese período de tono. Entonces, se ajusta su energía para que concuerde con la estimación del ruido de aspiración (ANE). Ya que se ha hallado informalmente que el uso de un filtro de conformación espectral no efectúa una diferencia perceptual, no se incluye en la paremetrización. La figura 15 muestra un diagrama de la aproximación de modelado del ruido de aspiración que se ha empleado.

1.3 Síntesis dentro del modelo de Síntesis de Análisis de Estimación Conjunta

La síntesis se efectúa siguiendo el Modelo JEAS de la figura 2 y aplicando los parámetros estimados durante el análisis. En teoría, cada marco K de la onda de habla, que se corresponde con un período de tono en segmentos sonoros y con un segmento fijo (en un ejemplo particular, un segmento fijo de 10 milisegundos) en partes no sonoras, puede generarse filtrando la señal estimada de excitación sonora y no sonora *e(n)* con el filtro del tracto vocal vt para ese marco particular

$$s_k(n) = e_k(n) * vt_k = e_k(n) - \sum_{i=1}^{p} a_i \cdot s_k(n-i), \quad n = 1 \cdots N_k \tag{17}$$

donde *p* es el orden del filtro y N_k es el número de muestras en el marco. La señal de excitación se construye bien mediante la adición de las estimaciones ajustadas del ruido LF y de aspiración, *1f(n)* y un (*n*), o bien generando simplemente una fuente de ruido gaussiano, *gn(n)*, en los segmentos sonoros (V) y no sonoros (U), respectivamente.

$$e_k(n) = \begin{cases} lf_k(n) + an_k(n), & k = V \\ gn_k(n), & k = U \end{cases}$$

(18)

en la práctica, ya que el análisis JEAS descrito se hace independientemente para cada marco, la continuidad de los parámetros estimados entre marcos adyacentes no está garantizada, particularmente dentro de los segmentos sonoros. Como resultado, se producen distorsiones preceptúales cuando los parámetros cambian demasiado bruscamente de marco a marco. Para reducir este problema, las trayectorias de los parámetros de la fuente glotal sonora y del tracto vocal se suavizan antes de la resíntesis.

Con respecto al tracto vocal, los coeficientes de filtro conjuntamente estimados ($\alpha_1 \ldots \alpha_p$) se convierten primero en Frecuencias Espectrales en Línea (LSF) debido a sus mejores propiedades de interpolación. Entonces, cada conjunto de los coeficientes LSF LSF^P ($lsf_1 \ldots lsf_P$) se promedian con los de los marcos anterior y siguiente para obtener una estimación de filtro de tracto vocal más suave para la síntesis

$$\mathrm{LSF}_k^p = \sum_{i=k-1}^{i=k+1} \mathrm{LSF}_i^p/3 \ .$$

(19)

Como para la fuente glotal, se sigue una aproximación similar. Primero, los parámetros T de LF ajustados (T_p, T_e, T_a, T_c) se convierten en parámetros R (R_g, R_k, R_a) que son más adecuados para la interpolación ya que están normalizados con respecto al período fundamental. De nuevo, para suavizar sus trayectorias, cada conjunto de parámetros R se promedia con los de los marcos anterior y siguiente. Las trayectorias de la ANE de la energía del ruido de la aspiración se suavizan de la misma forma

$$\mathcal{R}_k = \sum_{i=k-1}^{i=k+1} \mathcal{R}_i/3 \ .$$

(20)

Una vez que los parámetros fuente (R_g, R_k, R_a, *ANE*) han sido promediados, se usan para recalcular las ondas glotales derivadas de LF suavizadas *lf (n)* y las estimaciones del ruido de aspiración modulado en amplitud *an (n)* a utilizar como filtro de excitación *e (n)* para la resíntesis.

Para resintetizar la onda del habla, se emplea el esquema de superposición – adición de la ecuación 21

$$\tilde{s}(n) = \sum_{k=1}^{K} w_k(n - kN_{sc}^k) \cdot sc_k(n - kN_{sc}^k) \ ,$$

(21)

donde K es el número total de marcos W_k es la ventana de Hamming de forma que

$$w_k(n) = \begin{cases} 0.54 - 0.46\cos(2pi\frac{n}{N_{sc}^k}), & 0 \le n \le N_{sc}^k \\ 0, & \textit{en caso contrario} \end{cases}$$

(22)

y SC_k es la contribución sintética de longitud $M^k_{SC} = N_{k-1} + N_k$ generada por

$$sc_k(n) = e_k(n - N_{k-1}) - \sum_{i=1}^{p} \alpha_i \cdot sc_k(n-i) \quad n = 1 \cdots N_{sc}^k$$

(23)

de manera que el K-esimo marco de síntesis de las muestras N_k se obtenga como

$$\tilde{s}(n + kN_k) = w_{k-1}(n + N_k)sc_{k-1}(n + N_k) + w_k(n)sc_k(n)$$

(24)

La figura 16 muestra un diagrama esquemático del esquema de síntesis de superposición – adición empleado.

1.4 Modificación de tono y de escala de tiempo

Debido al modelado explícito e independiente del período fundamental y a las posibilidades de interpolación de las parametrizaciones del tracto vocal y de la fuente glotal empleadas, se implementan fácilmente las modificaciones de tono y de escala de tiempo dentro del marco JEAS.

Ambas transformaciones de tono y de escala de tiempo se basan en una aproximación de interpolación de trayectoria de parámetros, en la que la primera tarea comprende el cálculo del número de marcos requeridos en un segmento particular para conseguir las modificaciones deseadas. Una vez calculado el número modificado de marcos, los contornos del tamaño de los marcos, la excitación y las trayectorias de los parámetros del tracto vocal son nuevamente muestreados en el número modificado de marcos usando, por ejemplo, interpolación de ajustador cúbico. Ya que el modelo JEAS es síncrono en el tono, los tamaños de los marcos se corresponden con los períodos del tono en los segmentos sonoros mientras que son fijos en los segmentos no sonoros. Debido a sus mejores características de interpolación, se emplean coeficientes de LSF y parámetros R durante las transformaciones de tono y de escala de tiempo para representar respectivamente los parámetros del tracto vocal y de la fuente glotal, además de las energías de la ANE de aspiración y de ruido GNE gaussiano.

La modificación de escala de tiempo se lleva a cabo aumentando o disminuyendo el número de marcos por segmento e interpolando consecuentemente las pistas de los parámetros. Por ejemplo, para aumentar la duración de un segmento sonoro de *f* marcos en un 25%, el número modificado de marcos se calcula como *mf* = *f* + 0,25f. Entonces, se muestrea nuevamente el contorno del período del tono del punto *f* en el nuevo conjunto de puntos mf uniformemente separados, según se muestra en la figura 17. De esta forma, se conserva el contorno del período fundamental, es decir, la entonación, mientras se ralentiza su variación. Es necesario aplicar la misma repetición del muestreo a cada coeficiente LSF, parámetro R y pistas de ANE, para sintetizar el habla modificada en el tiempo. Los segmentos no sonoros también pueden escalarse en el tiempo usando el procedimiento descrito. En este caso, las trayectorias de los parámetros de excitación para el nuevo muestreo son las energías de GNE de la fuente de ruido gaussiano.

El tono puede alterarse multiplicando simplemente el contorno del período fundamental por un factor de escalado.

Por ejemplo, si un contorno de período de tono dado de *f* marcos $T = \{T^1, T^2,\ldots,T^f\}$ se multiplica por 0,5, el habla sintetizada con el contorno modificado $T = 0{,}5\ T = \{T^1, T^2,\ldots,T^f\}$ podría percibirse como que tiene el doble de la frecuencia fundamental original. Sin embargo, su duración también podría percibirse como la mitad del original. La escalado de los períodos fundamentales comprende la modificación de los tamaños de los marcos y, como consecuencia, las duraciones de los segmentos. Por esta razón, también es necesario modificar el número de marcos en un segmento cuando se escala el tono si se quiere mantener su duración. El número modificado de marcos mf en los períodos fundamentales escalados cuya duración se aproxima al original puede calcularse como

$$mf = f + \frac{(\bar{T} - \bar{T}') * f}{\bar{T}'}$$

(25)

donde $\bar{T}$ es el período fundamental medio original, $\bar{T}'$ es el período fundamental medio escalado. Una vez que se ha calculado mf, el contorno del período del tono escalado T´, los coeficientes de LSF, los parámetros R y las trayectorias de ANE deben muestrearse de nuevo en el nuevo número de marcos antes de resintetizar la onda del habla modificada en el tono.

2. Conversión de voz

En esta sección, se explora el uso de la parametrización de la fuente glotal JEAS y las transformaciones lineales probabilísticas continuas para la conversión de la voz y el rendimiento del Marco de Conversión de Voz JEAS resultante se compara con el de un sistema de VC Sinusoidal convencional (H. Ye y S. Young, High Quality Voice Morphing in Proc. ICASSP, volumen 1, pág. 1-9-12, 2004), referenciado como PSHM. La primera sección detalla el modelo del habla y las técnicas de transformación de características empleadas en la implementación de VC de JEAS. A continuación se presenta la medición objetiva de su cubierta espectral y el rendimiento de la conversión de la fuente de voz y la evaluación subjetiva de la capacidad de reconocimiento y la calidad de la salida convertida.

2.1 Conversión de Voz JEAS

A continuación se describen los procedimientos de transformación de la cubierta espectral y de la onda glotal empleados en la conversión de voz JEAS. Aunque la conversión de la cubierta espectral se hace de una forma similar a la implementación de conversión de voz sinusoidal ya conocida, la principal ventaja del modelo JEAS, es decir, la parametrización de la fuente de voz, permite que las características de la fuente se transformen también para coincidir con la diana. Al igual que ofrece el potencial para mejorar la fidelidad en la identidad de la diana, esto también evita la necesidad de procedimientos convencionales de predicción residual. Además, ya que la parametrización JEAS no implica una división de magnitud y fase del espectro, no se producen los defectos debidos a las discordancias de magnitud y fase convertidas y, así, no se requiere el uso de técnicas adicionales, tales como la predicción de fase.

2.1.1 Conversión de Cubierta Espectral

Los coeficientes de filtro de tracto vocal omnipolar de JEAS conjuntamente estimados ($\alpha_1 \ldots \alpha_p$) se convierten en parámetros de LSF mediante escalado de Bark para la transformación de las cubiertas espectrales JEAS. Primero, se calcula la respuesta de frecuencia lineal del filtro de tracto vocal conjuntamente estimado. Este se muestrea nuevamente de acuerdo con la escala de Bark, usando por ejemplo, la técnica de interpolación de ajustador cúbico ya conocida. Entonces se calculan los coeficientes de filtro omnipolar distorsionados aplicando, por ejemplo, el algoritmo convencional de Levinson – Durbin a la secuencia de autocorrelación del espectro de potencia distorsionado, entonces los coeficientes de filtro se transforman en LSF para la conversión.

Se emplea una función de transformación lineal probabilística continua para convertir las cubiertas espectrales LSF. Se usan Modelos de Mezclas Gaussianas (GMK) para describir los espacios vectoriales de las características glotales de la fuente y de la diana, clasificarlas en M clases y preparar las transformaciones lineales específicas para las clases. Entonces se emplea una suma ponderada de las transformaciones lineales para convertir cada vector x de características de la fuente glotal.

$$\mathcal{F}(x) = \left(\sum_{m=1}^{M} \lambda_m(x) W_m \right) \bar{x}$$

en la que $\bar{\bar{X}}$ es el vector de características extendido $\bar{\bar{X}} = [x' ; 1]'$ y _m es el peso de la interpolación de la matriz de transformación W_m, su valor viene dado por la probabilidad del vector x que pertenece a la clase C_m.

$$\lambda_m(x) = P(C_m|x) = \frac{\alpha_m N(y_t; \mu_m, \Sigma_m)}{\sum_{i=1}^{M} \alpha_i N(y_t; \mu_i, \Sigma_i)}$$

siendo α_m, μ_m y Σ_m los pesos, medias y varianzas de los componentes GMM respectivamente y N() representa la Distribución Normal. Las matrices de transformación W_m se estiman usando datos de entrenamiento paralelos de la fuente y de la diana y un criterio de error cuadrático mínimo.

Después de la conversión, los nuevos parámetros de LSF se transforman en coeficientes de filtro omnipolar y se muestrean nuevamente para la escala lineal antes de la síntesis. Ya que el uso de las transformaciones lineales ensanchan los formantes del habla convertida, se aplica un post filtro perceptual para estrechar los anchos de banda de los formantes, aumentar los valles espectrales y afilar los picos de los formantes.

La figura 18 ilustra el JEAS frente a las cubiertas espectrales PSHM, y en ella puede verse que las cubiertas PSHM capturan la inclinación espectral pero en JEAS se codifica mediante ondas glotales en su lugar. Además, aunque ambos

procedimientos están dirigidos a representar los formantes más importantes, existen pequeñas diferencias en sus amplitudes, frecuencias y/o anchos banda.

2.1.2 Conversión de la Onda Glotal

El trabajo previo sobre la conversión de la onda glotal ha demostrado que la cuantificación de los parámetros glotales es posible y capaz de capturar diferencias de la calidad de la fuente de voz. Por ejemplo, Childers y colaboradores (Glottal source modelling for voice conversión, Speech Communication, 16:127 – 138, 1995) crearon códigos de 32 entradas de parámetros polinómicos de fuente de voz a partir de frases producidas con diferentes calidades de voz y dirigidas a conseguir la conversión entre fonaciones modales, rotas, entrecortadas, ásperas, en falsete, susurrantes y roncas. Sin embargo, experimentos que implicaban transformaciones entre fonaciones más similares, es decir, diferentes hablantes modales o procedimientos de conversión alternativos todavía no se han explorado. Tampoco se ha investigado el uso de la parametrización glotal de LF.

La aproximación de mutación de la onda glotal adoptada en la conversión de voz de JEAS emplea Transformaciones Lineales Probabilísticas continuas para mapear parámetros LF glotales de diferentes hablantes modales masculinos y femeninos, que son los tipos de hablantes más comúnmente utilizados en las aplicaciones de conversión de voz.

Las transformaciones lineales probabilísticas continuas se han seleccionado para ser la aproximación más robusta y eficiente encontrada para convertir cubiertas espectrales. Las limitaciones de los procedimientos de conversión basados en códigos para transformaciones de cubiertas, es decir las discontinuidades provocadas por el uso de un número discreto de entradas de código, también pueden interpolarse para la modificación de ondas glotales. Así, el uso del modelo probabilístico continuo y de las transformaciones se espera que consiga también mejores conversiones glotales.

Los vectores de características empleados para convertir las características de la fuente glotal se derivan a partir de los parámetros del modelo JEAS enlazado con la fuente de voz de cada período de tono, es decir, la fuerza de excitación glotal E_e y los parámetros T ((T_p, T_e, T_a, T_c) obtenidos a partir del procedimiento de ajuste de LF y de la energía de la estimación del ruido de aspiración (ANE) usada para ajustar la del ruido gaussiano modulado en amplitud síncrono en tono modelado. Para normalizar los parámetros T dependientes de T_o para la conversión, se transforman en parámetros R (R_g, R_k, R_a), dando como resultado el vector de características de dimensión cinco (E_e, R_g, R_k, R_a, ANE) para la conversión de la onda glotal. Según se muestra en la figura 19, la aproximación de conversión glotal descrita es capaz de llevar los contornos de los parámetros del vector de característica de la fuente más cerca de la diana que, como consecuencia, también produce formas glotales convertidas más similares a la diana. En particular, la figura 19 muestra la transformación lineal de las Ondas Glotales LF: a) ondas glotales LF fuente, diana y derivada convertida; b) trayectorias fuente, diana y convertida de los parámetros del vector de características glotales (E_e, R_g, R_k, R_a, ANE).

2.2 Experimento: comparación entre un procedimiento de conversión de voz sinusoidal convencional y un procedimiento de Conversión de Voz JEAS

A continuación, se describe un experimento, en el cual un procedimiento de conversión de voz sinusoidal convencional (H. Ye y S. Young, High Quality Voice Morphing en Proc. ICASSP, volumen 1, pág. 1-9-12, 2004), denominado PSHM se compara con el rendimiento del procedimiento JEAS. Ambos procedimientos se han evaluado en una tarea de conversión basada en la base de datos VOICES (A. Kain, High Resolution voice transformation, thesis PhD , Oregon Health and Science University, 2001). Específicamente diseñado para propósitos de conversión de voz, el corpus se compone de tres ejemplos de 50 frases fonéticamente ricas habladas por 10 hablantes (5 hombres, 5 mujeres), es decir, un total de 150 palabras por hablante. Los datos del habla se grabaron usando una aproximación de “mimetización” que dio como resultado un alineamiento temporal natural entre frases idénticas producidas por los diferentes hablantes y sin tomar en consideración las referencias prosódicas de la identidad del hablante en alguna extensión. Para cada frase también se suministraron los instantes de cierre glotal derivados de las señales laringográficas y se han usado para el análisis síncrono de tono tanto de PSHM como de JEAS. Se han investigado cuatro experimentos diferentes de conversión de voz: transformaciones de hombre en hombre (MM), hombre en mujer (MF), mujer en hombre (FM) y mujer en mujer (FF). Las primeras 120 frases se utilizan para entrenamiento y las 30 restantes para comprobar cada conversión de par de hablantes.

A todo lo largo de los experimentos de conversión se han empleado vectores espectrales de LSF de orden 30 para entrenar 8 transformaciones de cubierta espectral entre cada par de hablantes fuente y diana usando datos de entrenamiento paralelos de VOICES. Este número se ha seleccionado por ser capaz de conseguir pequeñas relaciones de distorsión espectral al mismo tiempo que se generaliza para los datos de prueba. Se obtuvieron pares de vectores fuente – diana aplicando alineamiento forzado para marcar los límites subfónicos y usando Distorsión de Tiempo Dinámica para constreñir adicionalmente su alineamiento de tiempo. Para la predicción residual y de fase, se han construido GMM de diana y códigos de 40 clases y entradas. Finalmente, se han llevado también a cabo conversiones de onda glotal usando 8 transformaciones lineales por cada par de hablantes. Se han utilizado evaluaciones objetivas y subjetivas para comparar el rendimiento de los dos procedimientos.

2.2.1 Evaluación de Objetivos

2.2.1.1 Conversión de Cubierta Espectral

Ya que actualmente se aplican transformaciones de cubierta espectral lineal a los vectores de LSF, su rendimiento de conversión puede evaluarse fácilmente comparando las distancias de los vectores de LSF fuente, diana y convertido. Si la distancia entre dos vectores de LSF lsf_1 y lsf_2 se define como

$$D_{\mathrm{LSF}}(\mathrm{lsf}_1, \mathrm{lsf}_2) = |\mathrm{lsf}_1 - \mathrm{lsf}_2| = \sqrt{(\mathrm{lsf}_1 - \mathrm{lsf}_2)'(\mathrm{lsf}_1 - \mathrm{lsf}_2)} \ , \tag{26}$$

puede usarse la siguiente relación de distorsión R_{LSF} como una medida de objetivos de cómo cerrar los vectores fuente que se han convertido en la diana

$$R_{\mathrm{LSF}} = \frac{\sum_{t=1}^{L} D_{\mathrm{LSF}}(\mathrm{lsf}_{conv}(t), \mathrm{lsf}_{tgt}(t))}{\sum_{t=1}^{L} D_{\mathrm{LSF}}(\mathrm{lsf}_{src}(t), \mathrm{lsf}_{tgt}(t))} \cdot 100 \ , \tag{27}$$

donde lsf_{src}-(t), $lsf_{tgt}(t)$ y $lsf_{conv}(t)$ son respectivamente los vectores de LSF fuente, diana y convertido y se calcula la suma sobre los datos de prueba alineados en el tiempo, siendo L el número total de vectores de prueba después del alineamiento. Observe que una relación de distorsión del 100% se corresponde con la distorsión entre la fuente y la diana.

Se han calculado las relaciones R_{LSF} para las conversiones de cubierta espectral PSHM y JEAS sobre el conjunto de prueba de VOICES. La figura 20 muestra los resultados obtenidos. Aunque las diferencias son pequeñas, se ha encontrado que JEAS funciona ligeramente mejor que PSHM con relaciones de distorsión de LSF de un 3% menor en todas las tareas de conversión. Esto podría deberse al hecho de que las cubiertas espectrales de JEAS no codifican información de inclinación espectral, lo que reduce las variaciones de LSF provocadas por las diferencias de inclinación dando como resultado mapeos más exactos.

2.2.1.2 Conversión de Fuente de Voz

También pueden utilizarse medidas similares de distorsión de objetivos par evaluar la conversión de las características de la fuente de voz, es decir, Predicción Residual y Conversión de Onda Glotal en implementaciones PSHM y JEAS, respectivamente.

La Predicción Residual reintroduce los detalles espectrales de la diana no capturados por la conversión de la cubierta espectral, dando como resultado unos espectros convertidos del habla más cercanos a la diana. La Conversión de Onda Glotal, por otra parte, mapea representaciones del dominio de tiempo de las ondas glotales lo que en el dominio de frecuencia da como resultado una mejor concordancia de los formantes glotales y las inclinaciones espectrales de los espectros convertidos. Aunque los procedimientos son diferentes, su efecto espectral es similar, es decir se dirigen a reducir las diferencias entre los espectros del habla convertida y diana.

Una forma de evaluar si los procedimientos de conversión de la fuente de voz consiguen el efecto deseado es medir las distancias espectrales logarítmicas (LSD) entre los espectros convertido y diana antes y después de la conversión de la fuente de voz. Si la distancia espectral logarítmica RMS entre dos espectros se define como

$$D_{\mathrm{LSD}}(S_1, S_2) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(10log_{10}(\mathrm{amp}_k^1) - 10log_{10}(\mathrm{amp}_k^2)\right)^2} \tag{28}$$

donde $\{amp_k\}$ son amplitudes armónicas nuevamente muestreadas a partir del espectro S en los puntos K sobre la escala de frecuencias de Bark (K se ha fijado en 100 puntos en este trabajo). Entonces, puede usarse una relación de distorsión R_{LSD} similar a R_{LSF} para comparar las distancias espectrales logarítmicas convertida – diana con y sin conversión de la

fuente de voz.

$$R_{\mathrm{LSD}} = \frac{\sum_{t=1}^{L} D_{\mathrm{LSD}}(S_{conv}(t), S_{tgt}(t))}{\sum_{t=1}^{L} D_{\mathrm{LSD}}(S_{orig}(t), S_{tgt}(t))} \cdot 100$$

(29)

En la que $S_{conv}(t)$ y $S_{orig}(t)$ son los espectros convertidos respectivamente con y sin conversión de fuente de voz y $S_{tgt}(t)$ es el espectro de la diana. Así, una relación del 100% se corresponde con la distorsión entre los espectros convertidos de la cubierta espectral sin transformación de la fuente de voz y los espectros diana.

La figura 21 ilustra las relaciones R_{LSD} calculadas para Predicción Residual y Conversión de Onda Glotal sobre el conjunto de prueba. Los resultados muestran que ambas técnicas de conversión de fuente de voz reducen las distorsiones entre los espectros del habla convertido y diana. La Predicción Residual funciona ligeramente mejor, principalmente a causa de que el algoritmo está diseñado para producir residuos que minimizan la distancia espectral logarítmica representada en R_{LSD}. En contraste, la conversión de onda glotal está dirigida a minimizar el error de conversión de los parámetros glotales sobre los datos de entrenamiento y no la distancia espectral logarítmica. No obstante, ambos procedimientos tienen éxito al llevar los espectros convertidos cerca de la diana.

2.2.2 Evaluación Subjetiva

Para comparar perceptualmente los sistemas de conversión de voz PSHM y JEAS, se llevó a cabo un ensayo de audición para comprobar su rendimiento en términos de capacidad de reconocimiento y calidad. 12 sujetos tomaron parte en el estudio perceptual que constó de dos partes.

La primera parte fue una prueba ABX en la cual se presentaron a los sujetos palabras convertidas con PSHM (A), convertidas con JEAS (B) y diana (X) y se les pidió que seleccionaran la muestra A o B que ellos encontraban más similar a la diana X en términos de identidad del hablante. Las características de las cubiertas espectrales y de la fuente de voz se transformaron con los procedimientos descritos anteriormente para cada sistema, es decir se usaron predicción de conversión, de residuo y de fase de cubierta espectral para transformaciones PSHM y conversión de cubierta espectral y de onda glotal para transformaciones JEAS. Además, se empleó la prosodia de la diana para sintetizar las frases convertidas para normalizar las diferencias de tono, duración y energía entre los hablantes fuente y diana para la comparación perceptual. Se presentaron 10 palabras de cada tipo de conversión (MM, MF, FM, FF). El orden de las muestras en términos de tipo de conversión y sistema de conversión fue aleatorio. Una audición informal de las palabras transformadas usando los sistemas de conversión PSHM y JEAS reveló que a menudo fue muy difícil escoger de forma convincente entre los sistemas en términos de identidad del hablante. Por esta razón, también se permitió a los sujetos seleccionar una opción “SIN PREFERENCIA CLARA” cuando encontraron difícil la selección o no tenían una preferencia clara hacia una de las muestras de habla A o B presentadas.

La figura 22 muestra los resultados de la prueba ABX. En todos los tipos de conversión, se prefirieron las muestras convertidas con JEAS en vez de las convertidas con PSHM, pero la diferencia de preferencia varía dependiendo del tipo de conversión, siendo por ejemplo casi la misma para las transformaciones FM. Sin embargo, la opción “SIN PREFERENCIA CLARA” (NSP) se ha seleccionado casi tan a menudo como las palabras convertidas con JEAS en general, lo que revela que los sujetos hallaron realmente difícil distinguir entre los sistemas de conversión en términos de identidad del hablante. Ya que las señales más importantes de identificación del hablante, es decir las cubiertas espectrales, se transforman usando el mismo procedimiento en las dos implementaciones de conversión, se espera que ambos sistemas funcionen igual en términos de capacidad de reconocimiento del hablante. Además, los resultados obtenidos muestran que las técnicas de Predicción Residual y Conversión de Onda Glotal también son comparables en términos de transformación de la identidad perceptual del hablante.

El segundo ensayo de audición se dirigía a determinar qué sistema produce un habla con una mayor calidad. Se presentó a los sujetos pares de sonidos del habla convertidos con PSHM y JEAS y se les pidió que seleccionaran el que pensaban que tenían una mejor calidad de habla. Los resultados se ilustran en la figura 23. Hay una preferencia clara por las palabras convertidas usando el procedimiento JEAS, seleccionado un 75,7% de media, lo que se desprendía de la diferencia de calidad claramente distinguible entre las muestras transformadas con PSHM y JEAS. Los sonidos obtenidos después de la conversión con PSHM tenían una calidad “ruidosa” provocada por las discontinuidades de fase que todavía existen a pesar de la Predicción de Fase. Comparativamente, las palabras convertidas con JEAS suenan más fluidas. Esta diferencia de calidad se piensa que también desvió ligeramente la preferencia para la conversión con JEAS en el ensayo ABX.

Entre otros, el procedimiento y el dispositivo de conversión de voz de la presente invención es aplicable a marcos que requieran transformaciones de claridad de voz. Como una de tales aplicaciones, puede mencionarse su uso para reparar las características de la fuente de voz anormal del habla traqueo-esofágica.

La invención obviamente no está limitada a las realizaciones específicas aquí descritas, sino que comprende cualquier variación que pueda ser considerada por cualquier persona experta en la materia (por ejemplo, con respecto a la selección de componentes, configuración, etc), dentro del alcance general de la invención tal como se define en las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un procedimiento para convertir una señal de habla de un hablante fuente en un a señal de voz convertida, que comprende los pasos de:

- una etapa de entrenamiento, en la cual:

- dada una base de datos de entrenamiento de datos fuente y diana paralelos, para cada período de tono de dicha base de datos de entrenamiento:

- modelar cada período de tono por medio de una onda glotal y un filtro de tracto glotal de acuerdo con el modelo de Lu y Smith, para obtener un conjunto de parámetros Liljencrants – Fant LF, dicho conjunto de parámetros LF comprende un parámetro de fuerza de excitación E_e y un conjunto de parámetros T, T_p, T_e, T_a, T_c que modelan una onda glotal y un conjunto de coeficientes de filtro de tracto vocal omnipolar ($\alpha_1 \ldots \alpha_p$);

- convertir dichos parámetros T, T_p, T_e, T_a, T_c en parámetros R_g, R_k, R_a;

- convertir dichos coeficientes de filtro de tracto vocal omnipolar ($\alpha_1 \ldots \alpha_p$) en frecuencias espectrales en línea en la escala de Bark $lsf_1 \ldots lsf_p$;

- definir un vector glotal G a convertir;

- definir un vector de tracto vocal de LSF a convertir, dicho vector de tracto vocal de LSF comprende dichas frecuencias espectrales en línea en la escala de Bark $lsf_1 \ldots lsf_p$;

- aplicar la eliminación del ruido de la ondícula para obtener una obtener una estimación del ruido de aspiración glotal;

- a partir del conjunto de vectores de tracto vocal de LSF obtenidos para cada período de tono de dicha base de datos de entrenamiento, estimar una función de transformación lineal probabilística continua del tracto vocal usando el criterio de error cuadrático mínimo;

el procedimiento se caracteriza porque dicha etapa de modelado comprende además los pasos de:

- modelar dicha estimación del ruido de aspiración modulando el ruido gaussiano de la varianza unitaria próxima a cero con la mencionada onda glotal modelada y ajustando su energía ANE para coincidir con la mencionada estimación del ruido de aspiración;

dicho vector glotal G a convertir comprende dicho parámetro de fuerza de excitación E_e, dichos parámetros R, R_g, R_k, R_a y dicha energía ANE de la estimación del ruido de aspiración,

el procedimiento comprende además:

- una etapa de conversión en la cual una onda de habla de prueba dada se modela y se transforma en un conjunto de parámetros E_e', R_g', R_k, R_a', ANE´, LSF´;

- una etapa de síntesis en la cual una onda de habla convertida se sintetiza a partir de dicho conjunto de parámetros convertidos E_e', R_g', R_k, R_a', ANE´, LSF´.

2. El procedimiento de acuerdo con la reivindicación 1, en el que dicha etapa de entrenamiento comprende además:

- a partir del conjunto de vectores glotales G obtenidos para cada período de tono de dicha base de datos de entrenamiento, estimar una función de transformación lineal probabilística continua de la onda glotal usando el criterio de error cuadrático mínimo.

3. El procedimiento de acuerdo con la reivindicación bien 1 ó bien 2, en el que dicho paso de modelar cada período de tono por medio de una onda glotal y un filtro de tracto vocal de acuerdo con el modelo de Lu y Smith, comprende los pasos de:

- modelar la onda glotal usando el modelo de Rosenberg – Klatt;

- usar optimización convexa para obtener un conjunto de parámetros glotales de Rosenberg – Klatt y los coeficientes del filtro de tracto vocal omnipolar $\alpha_1 \ldots \alpha_p$ en el que dicho paso de usar optimización convexa comprende un paso de pre-énfasis adaptativo para estimar y eliminar la contribución de filtro de inclinación espectral de la onda del habla antes de la optimización convexa.

4. El procedimiento de acuerdo con la reivindicación 3, en el que dicho paso de modelar cada período de tono por medio de una onda glotal y un filtro de tracto vocal de acuerdo con el modelo de Lu y Smith, comprende además los pasos de:

- obtener una onda glotal derivada mediante el filtrado inverso de dicho período de tono usando dichos coeficientes de filtro de tracto vocal omnipolar $\alpha_1 \ldots \alpha_p$;

- ajustar dicho conjunto de parámetros LF a dicha onda glotal derivada de filtrado inverso mediante estimación directa y optimización no lineal constreñida.

5. El procedimiento de acuerdo con cualquier reivindicación precedente, en el que dicha etapa de conversión comprende, para cada período de tono de dicha onda de habla de prueba:

- obtener un vector glotal G a convertir, comprendiendo dicho vector glotal un parámetro de fuerza de excitación E_e, un conjunto de parámetros R, R_g, R_k, R_a y la energía ANE de dicha estimación de ruido de aspiración;

- obtener el vector de tracto vocal de LSF a convertir, dicho vector de tracto vocal de LSF comprende un conjunto de frecuencias espectrales en línea en la escala de Bark $lsf_1 \ldots lsf_p$;

- aplicar dicha función de transformación lineal probabilística continua del tracto vocal estimada durante la etapa de entrenamiento para obtener un vector de parámetros de LSF de tracto vocal convertidos;

- transformar dicho vector glotal G usando dicha función de transformación lineal probabilística continua de onda glotal estimada durante la etapa de entrenamiento, obteniendo así un vector glotal convertido G´ que comprende un conjunto de parámetros convertidos E_e´, R_g´, R_k, R_a´, ANE´, LSF´;

6. El procedimiento de acuerdo con la reivindicación 5, en el que dichas etapas de obtener un vector glotal G a convertir y un vector de tracto vocal de LSF a convertir comprende además los pasos de:

- modelar cada período de tono por medio de una onda glotal y un filtro de tracto vocal de acuerdo con el modelo de Lu y Smith, para obtener un conjunto de parámetros LF, comprendiendo dicho conjunto de parámetros LF un parámetro de fuerza de excitación E_e y un conjunto de parámetros T, T_p, T_e, T_a, T_c que modelan la onda glotal, y un conjunto de coeficientes de filtro de tracto vocal omnipolar $\alpha_1 \ldots \alpha_p$;

- convertir dichos coeficientes de filtro de tracto vocal omnipolar en frecuencias espectrales en línea en la escala de Bark $lsf_1 \ldots lsf_p$;

- convertir dichos parámetros T en parámetros R, R_g, R_k, R_a;

- definir un vector glotal G a convertir;

- definir un vector de tracto vocal de LSF a convertir.

7. El procedimiento de acuerdo con la reivindicación bien 5 o bien 6, en el que dicha etapa de conversión comprende además un paso de post-filtrado de dicho vector de parámetros convertidos de tracto vocal de LSF´.

8. El procedimiento de acuerdo con cualquier reivindicación precedente, en el que dicha etapa de síntesis, en la cual se sintetiza dicha onda de habla convertida a partir de dicho conjunto de parámetros convertidos Ee´, Rg´, Rk, Ra´, ANE´, LSF´; comprende los pasos de:

- interpolar las trayectorias de dichos parámetros convertidos R_g´, R_k, R_a´, ANE´, LSF´; de cada período de tono, obteniendo así un conjunto de parámetros interpolados R_g", R_k", R_a", ANE", LSF"; que comprende los parámetros R_g", R_k", R_a", energía interpolada (ANE) y un vector de tracto vocal interpolado de LSF";

- convertir dicho vector de tracto vocal interpolado de LSF" en un vector de coeficientes de filtro omnipolar A";

- convertir dichos parámetros R interpolados R_g", R_k", R_a" en parámetros T interpolados T_p", T_e", T_a", T_c" ;

- para cada marco de dicha onda de voz de prueba, generar una señal de excitación $e_k(n)$, en la que k indica el k-enésimo marco.

9. El procedimiento de acuerdo con la reivindicación 8, en el que dicha etapa de generar una señal de excitación comprende, para cada uno de dichos marcos:

- si dicho marco es sonoro:

- a partir de dichos parámetros T interpolados T_p", T_e", T_a", T_c" y de dicho parámetro de fuerza de excitación E_e

generar una onda glotal interpolada $lf_k(n)$;

- a partir de dicho parámetro de energía interpolado ANE", generar un ruido de aspiración interpolado $an_k(n)$;

- generar dicha señal de excitación sonora $e_k(n)$ añadiendo dicha onda glotal interpolada $lf_k(n)$ y dicho ruido de aspiración interpolado $an_k(n)$;

- si dicho marco no es sonoro:

- generar dicha señal de excitación no sonora $e_k(n)$ a partir de una fuente de ruido gaussiano $gn_k(n)$.

10. El procedimiento de acuerdo con la reivindicación bien 8 o bien 9 en el que dicha etapa de síntesis comprende además:

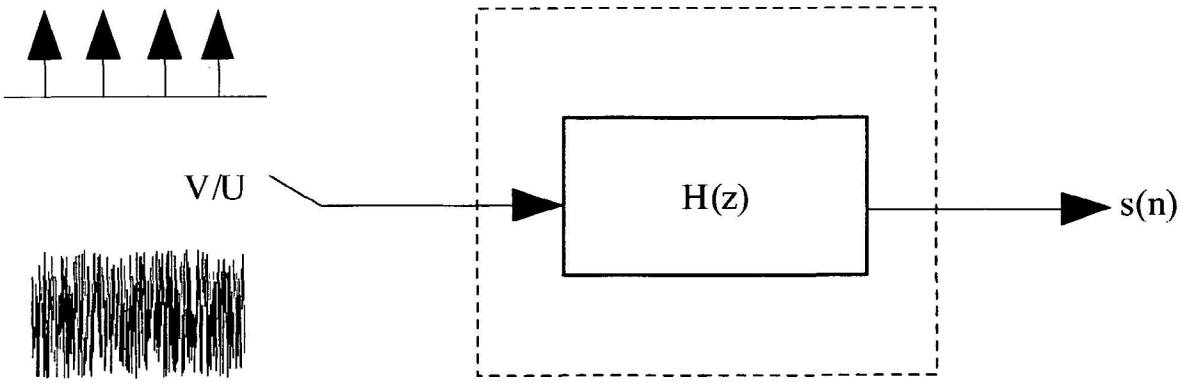
- generar una contribución sintética de cada marco filtrando dicha señal de excitación $e_k(n)$ con dicho vector de coeficientes de filtro omnipolar A";

- multiplicar dicha contribución sintética por una ventana de Hamming, superponer y añadir, para generar la señal de habla convertida.

11. Un procedimiento aplicable a transformaciones de calidad de voz, tales como la reparación del habla traqueo-esofágica, que comprende los pasos de procedimiento de cualquier reivindicación precedente.

12. Un dispositivo que comprende medios adaptados para llevar a cabo los pasos del procedimiento de cualquier reivindicación precedente.

13. Un medio de código de programa informático adaptado para realizar los pasos del procedimiento de cualquier reivindicación 1 – 11, cuando dicho programa se ejecuta en un ordenador, un procesador de señales digitales, una matriz de puertas programables por campo, un circuito integrado específico para la aplicación, un microprocesador, un microcontrolador o cualquier otra forma de hardware programable.



Onda glotal, tracto glotal y radiación de los labios

FIG. 1
(Técnica anterior)

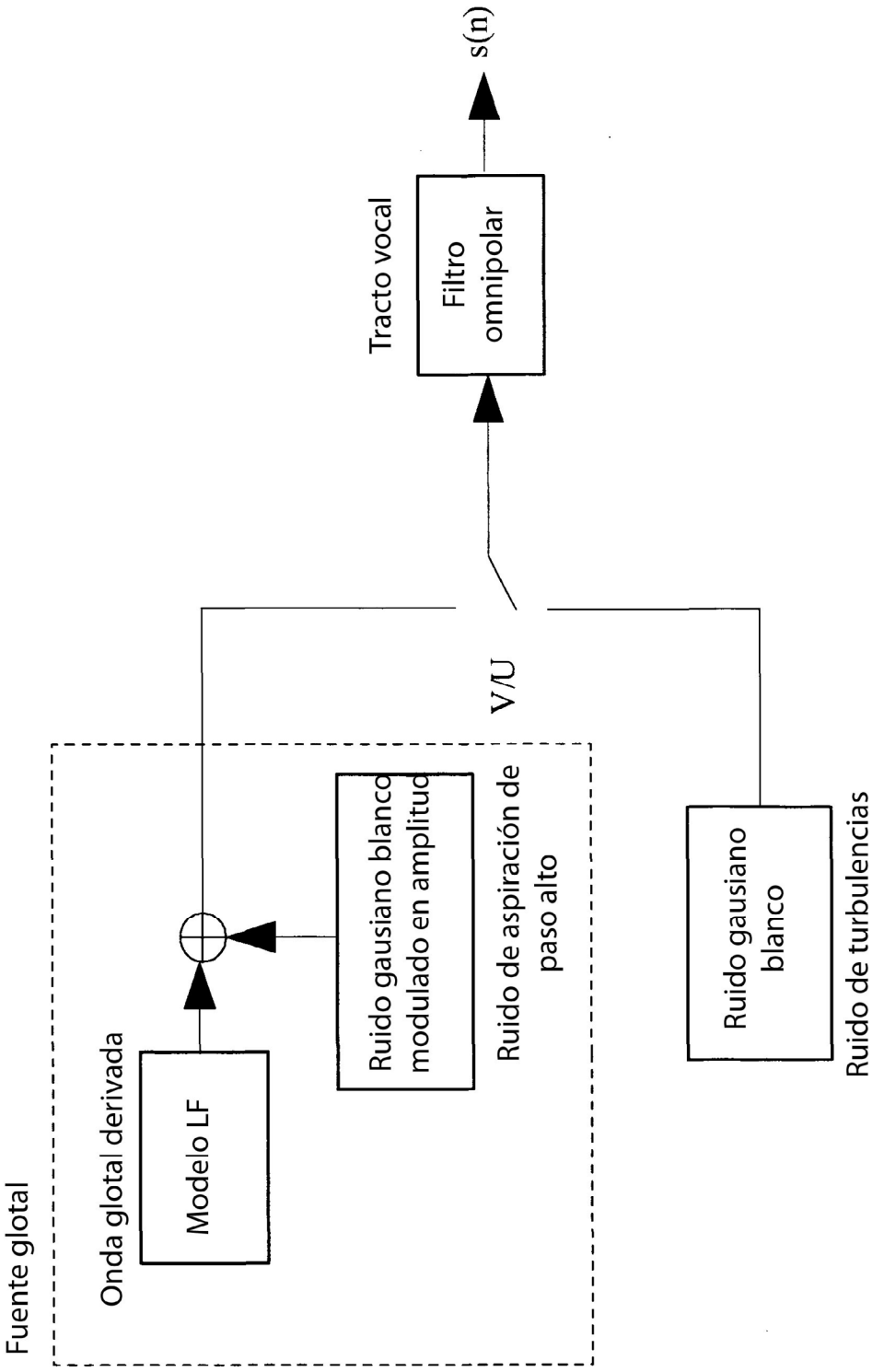

s(n)
Tracto vocal
Filtro omnipolar
V/U
Fuente glotal
Onda glotal derivada
Modelo LF
Ruido gausiano blanco modulado en amplitud
Ruido de aspiración de paso alto
Ruido gausiano blanco
Ruido de turbulencias

FIG. 2

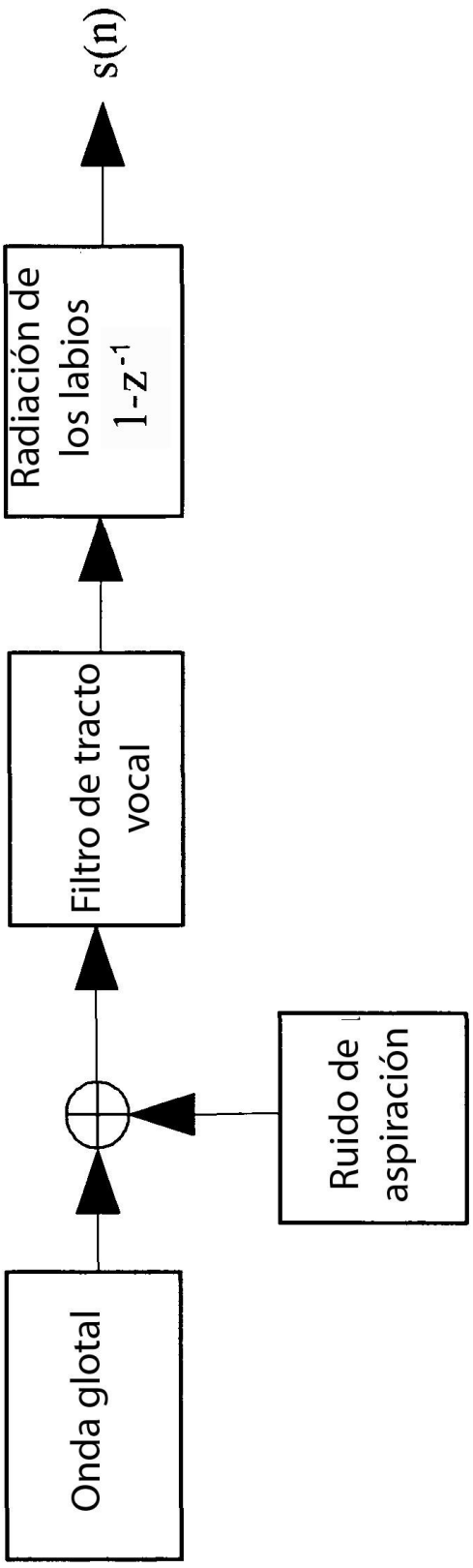

s(n)
Radiación de los labios
1-z-1
Filtro de tracto vocal
Ruido de aspiración
Onda glotal

FIG. 3

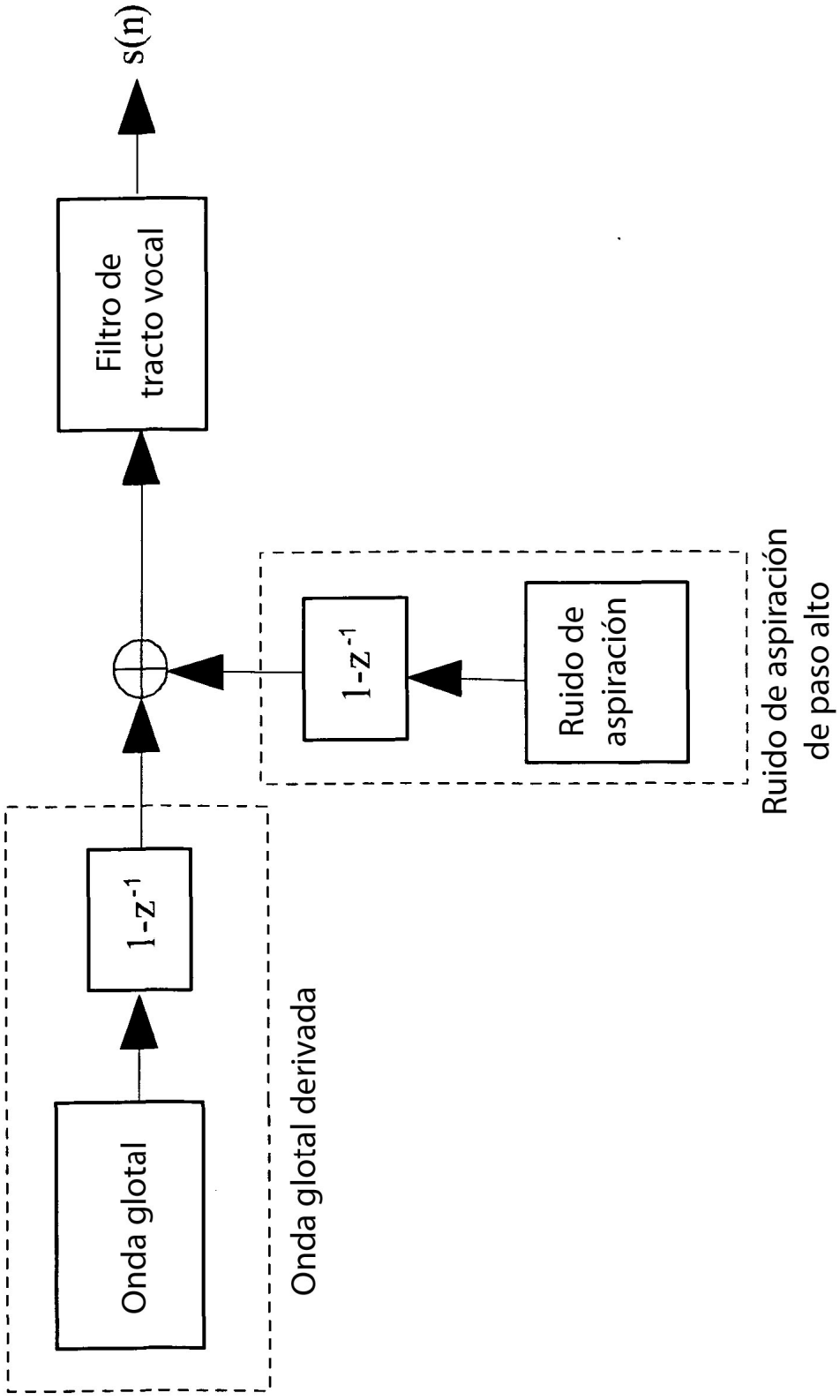

s(n)
Filtro de tracto vocal
Onda glotal
1-z-1
Onda glotal derivada
1-z-1
Ruido de aspiración
Ruido de aspiración de paso alto

FIG. 4

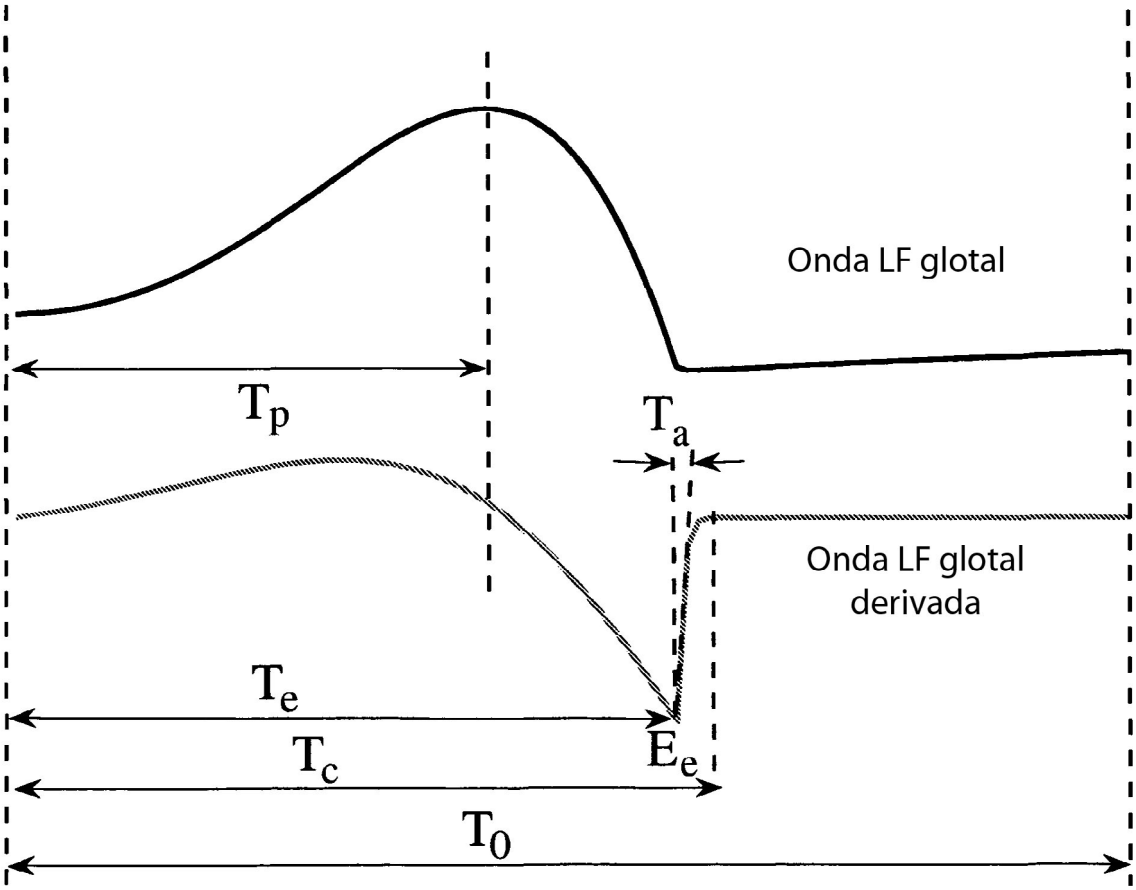


FIG. 5

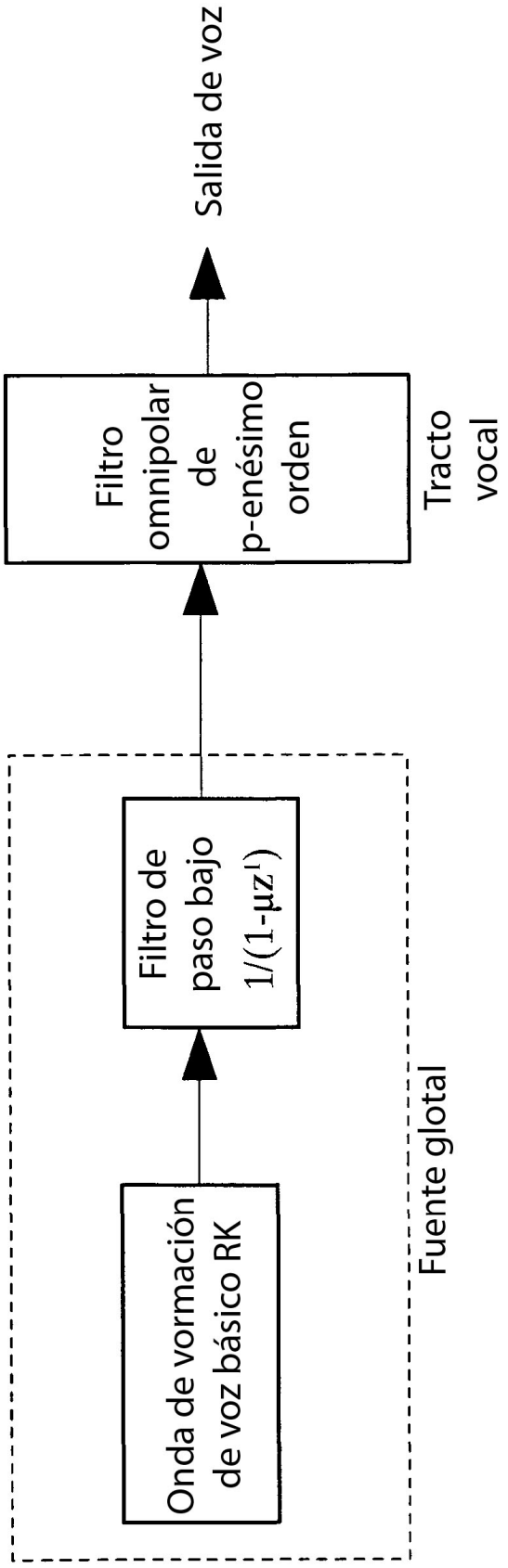

Salida de voz
Filtro omnipolar de p-enésimo orden
Tracto vocal
Filtro de paso bajo 1/(1-μz^-1)
Onda de vormación de voz básico RK
Fuente glotal

FIG. 6

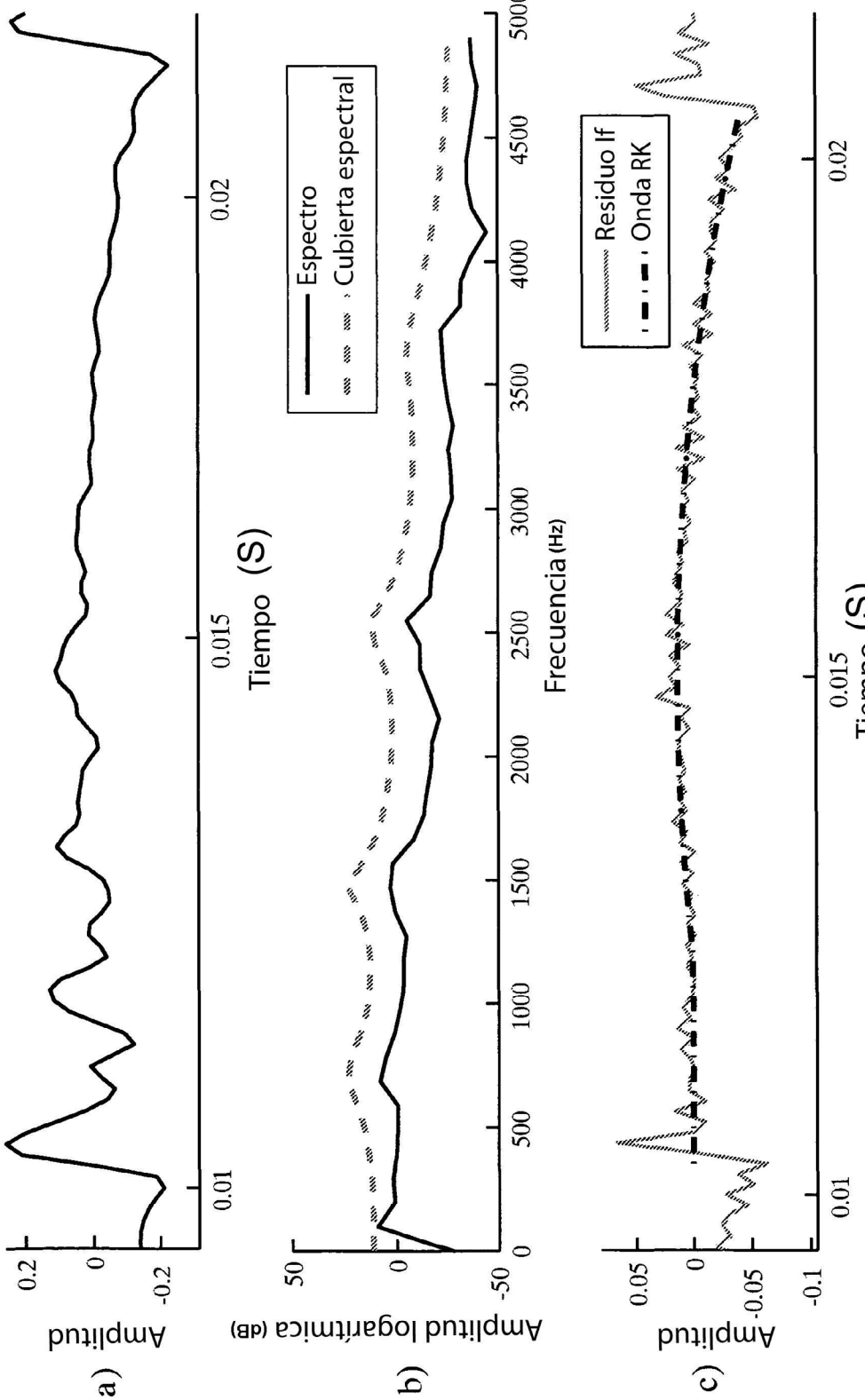

a)
Amplitud
0.2
0
-0.2
0.01
0.015
0.02
Tiempo (S)
b)
Amplitud logarítmica (dB)
50
0
-50
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Frecuencia (Hz)
Espectro
Cubierta espectral
c)
Amplitud
0.05
0
-0.05
-0.1
0.01
0.015
0.02
Tiempo (S)
Residuo lf
Onda RK

FIG. 7

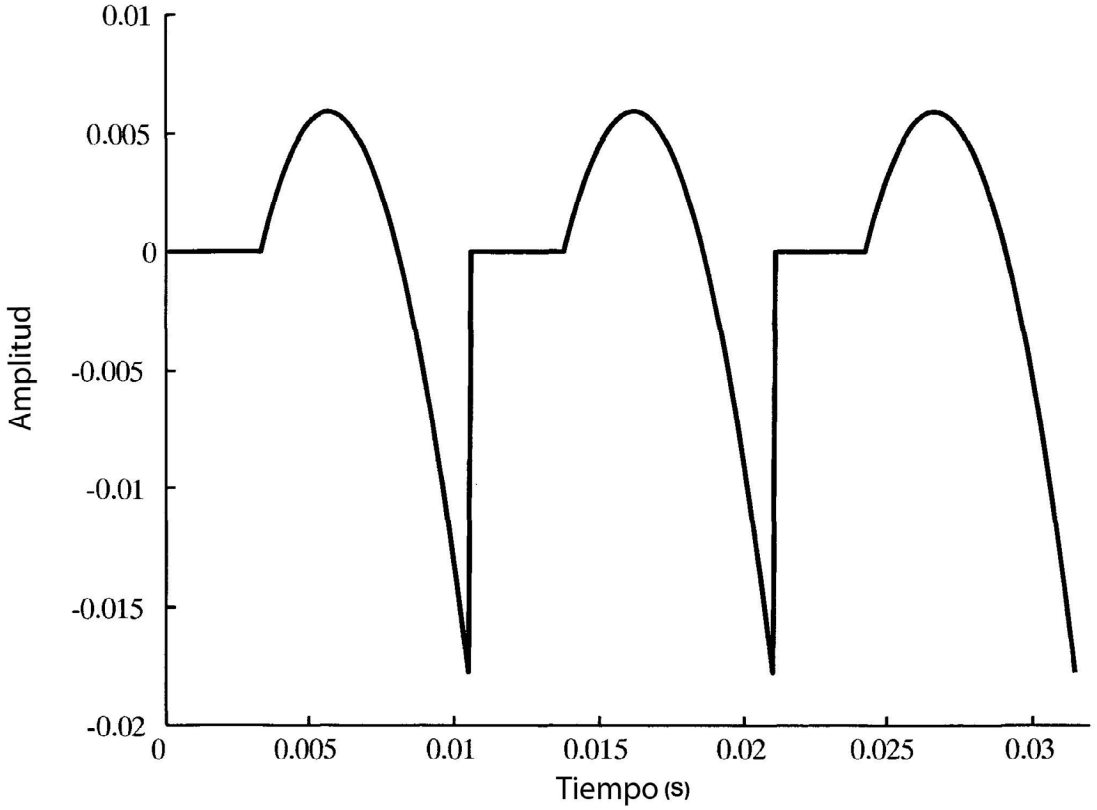

0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
Amplitud
0
0.005
0.01
0.015
0.02
0.025
0.03
Tiempo (s)

FIG. 8

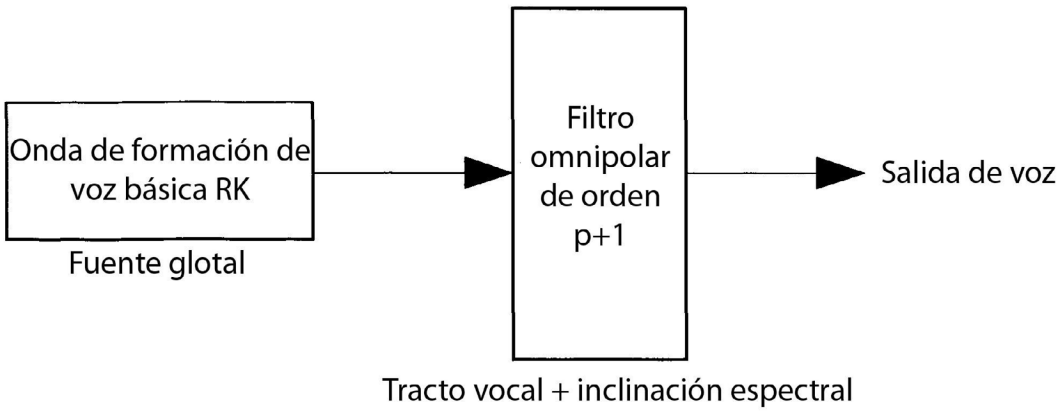

Onda de formación de voz básica RK
Fuente glotal
Filtro omnipolar de orden p+1
Salida de voz
Tracto vocal + inclinación espectral

FIG. 9

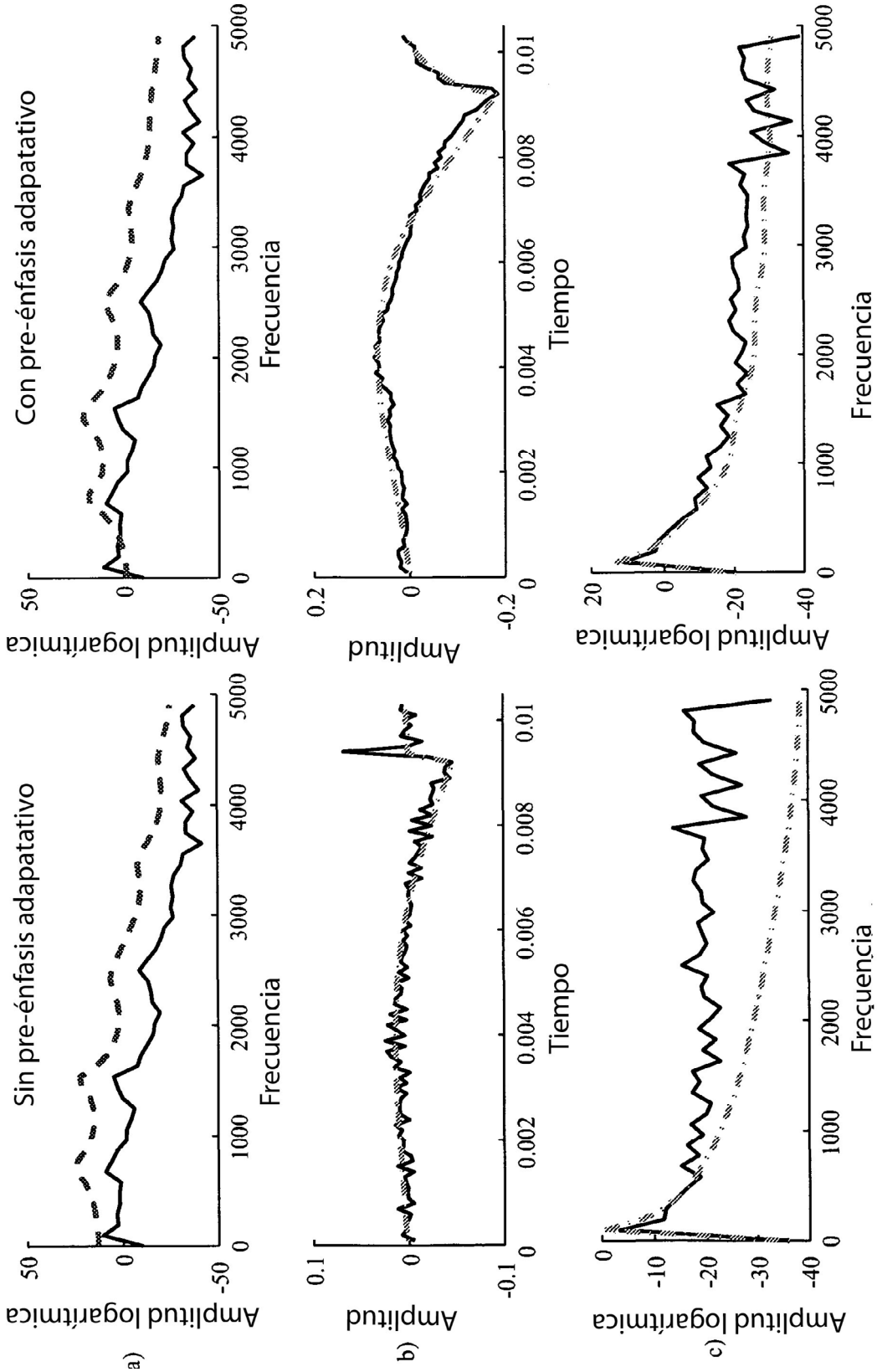

Sin pre-énfasis adapatativo
Con pre-énfasis adapatativo
Amplitud logarítmica
Frecuencia
Amplitud
Tiempo
a)
b)
c)

FIG. 10

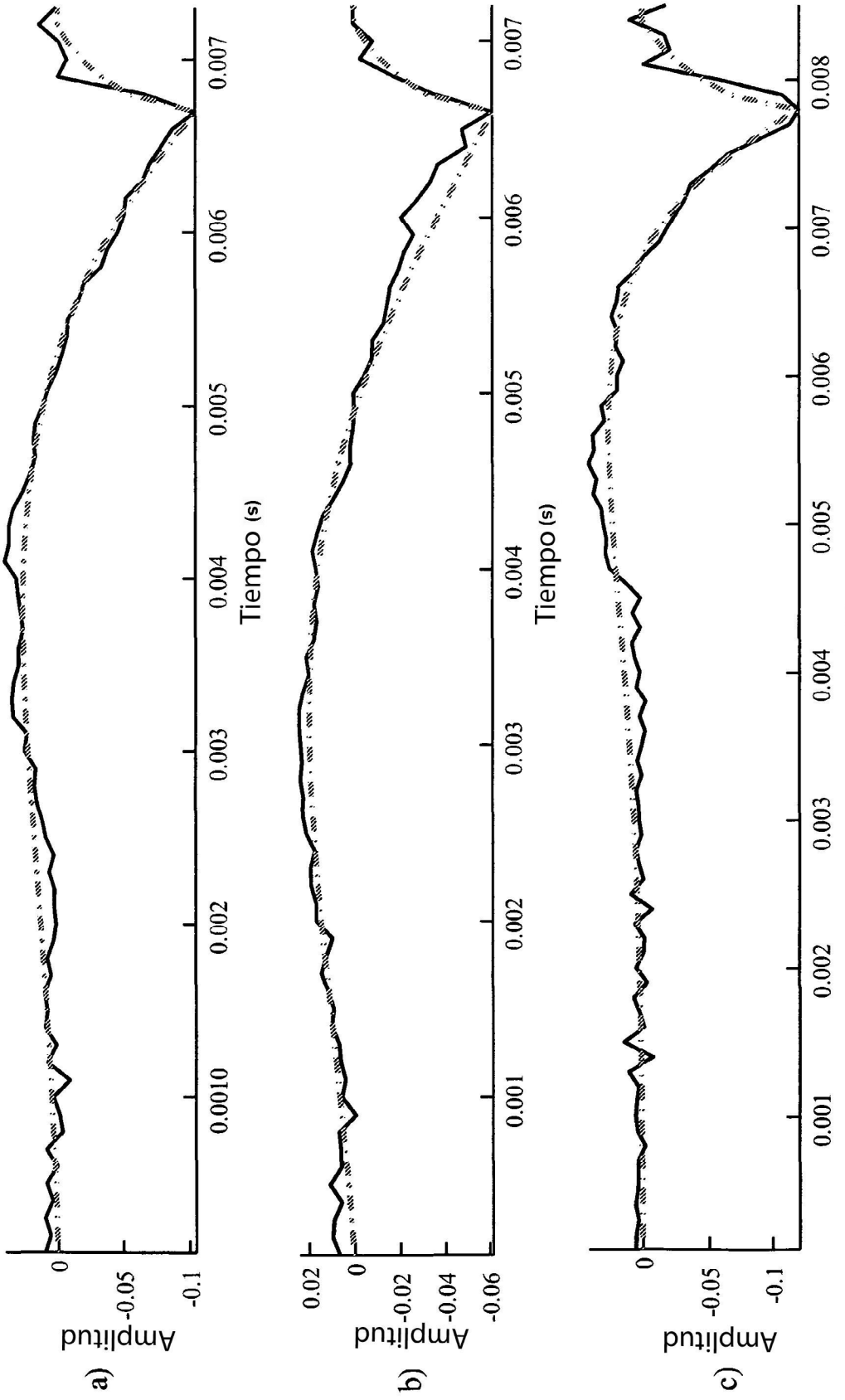

a)
Amplitud
0
-0.05
-0.1
0.0010
0.002
0.003
0.004
0.005
0.006
0.007
Tiempo (s)
b)
Amplitud
0.02
0
-0.02
-0.04
-0.06
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Tiempo (s)
c)
Amplitud
0
-0.05
-0.1
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
Tiempo (s)

FIG. 11

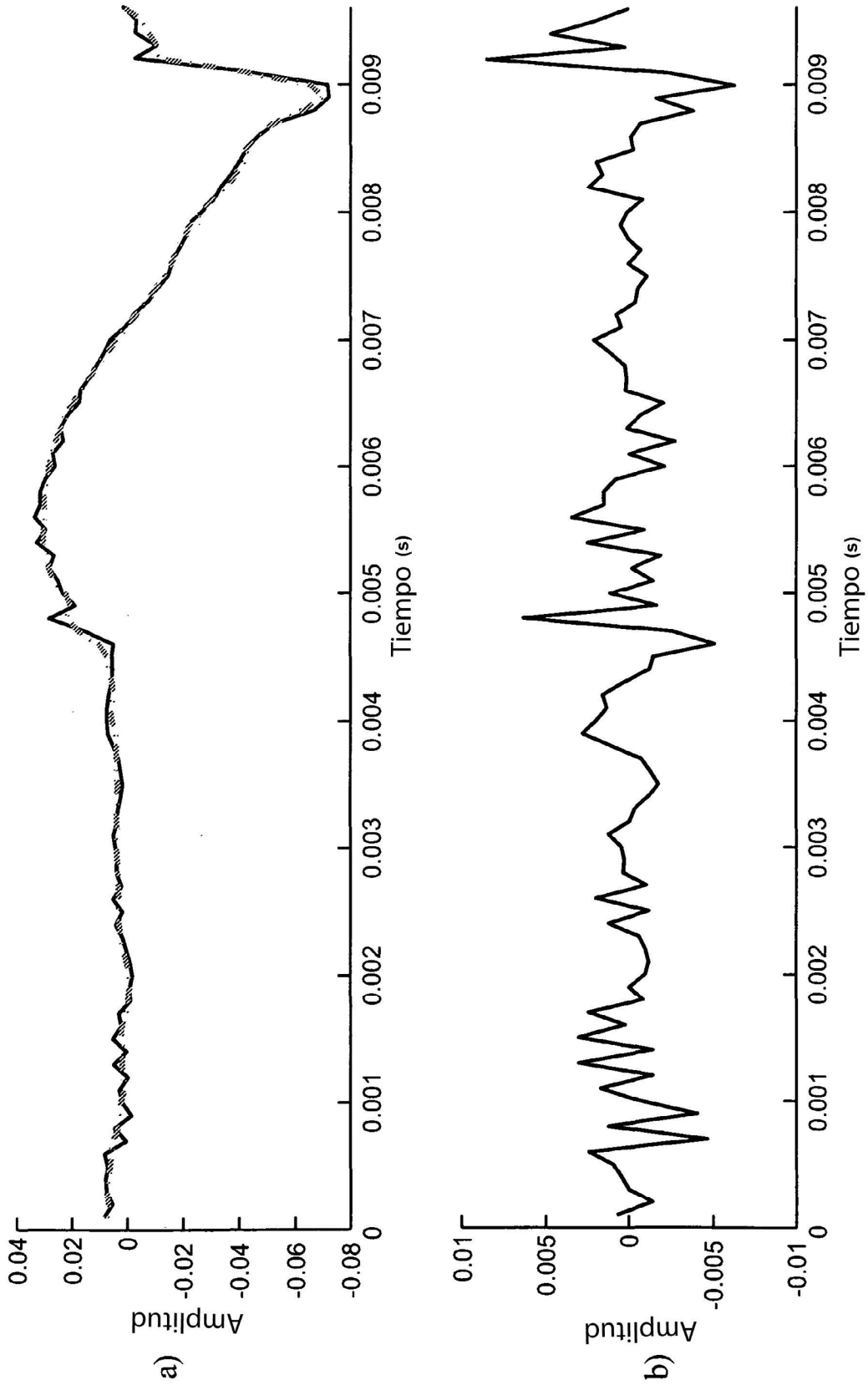

a)
Amplitud
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
Tiempo (s)
b)
Amplitud
0.01
0.005
0
-0.005
-0.01
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
Tiempo (s)

FIG. 12

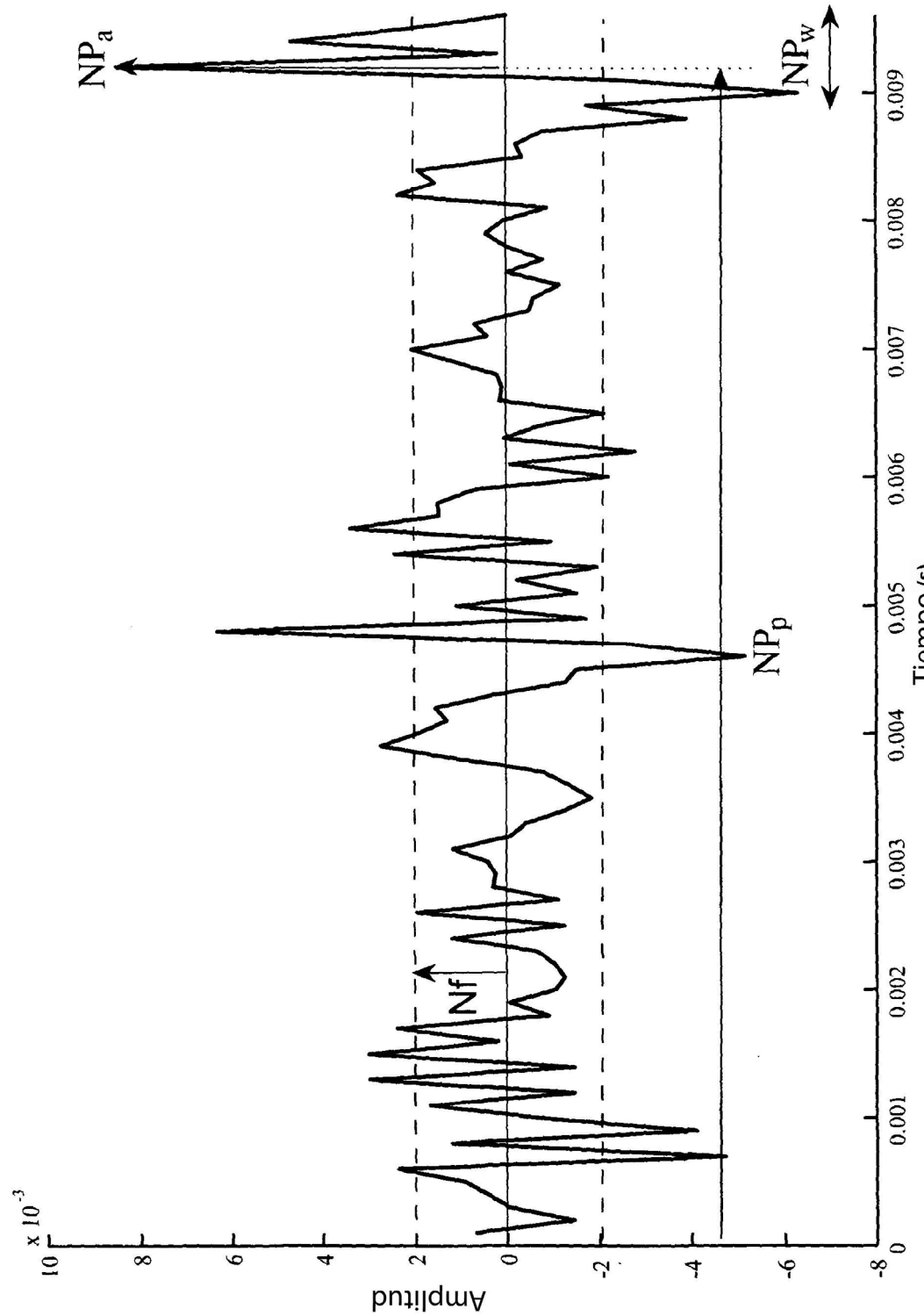

NPa
NPw
NPp
Nf
Amplitud
Tiempo (s)
x 10-3
10
8
6
4
2
0
-2
-4
-6
-8
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009

FIG. 13

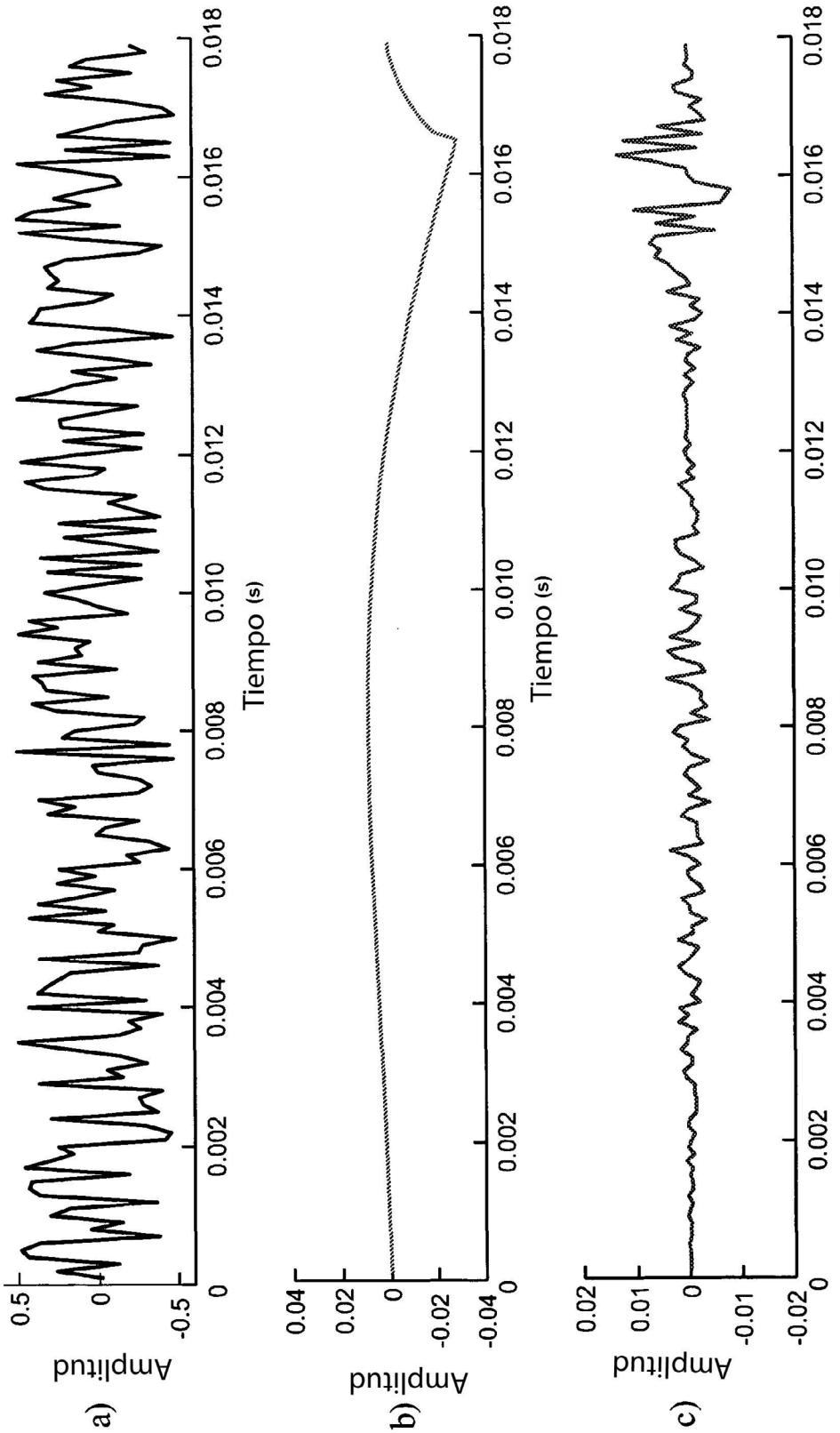

a)
Amplitud
0.5
0
-0.5
0
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0.018
Tiempo (s)
b)
Amplitud
0.04
0.02
0
-0.02
-0.04
0
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0.018
Tiempo (s)
c)
Amplitud
0.02
0.01
0
-0.01
-0.02
0
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0.018
Tiempo (s)

FIG. 14

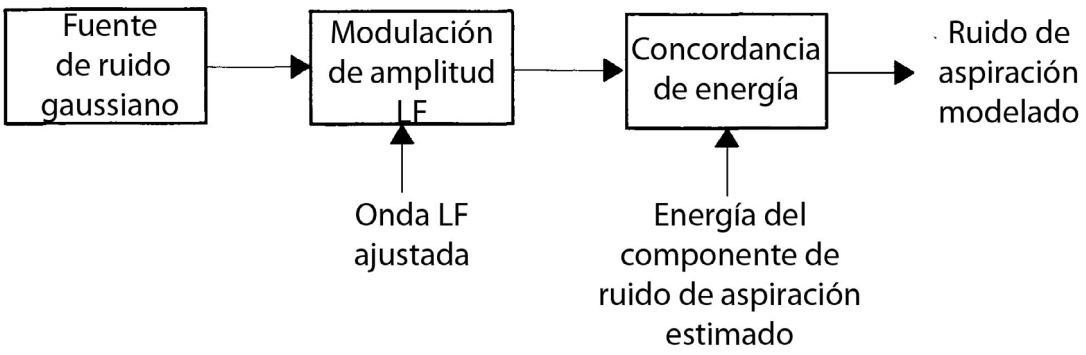

Fuente de ruido gaussiano
Modulación de amplitud LF
Concordancia de energía
Ruido de aspiración modelado
Onda LF ajustada
Energía del componente de ruido de aspiración estimado

FIG. 15

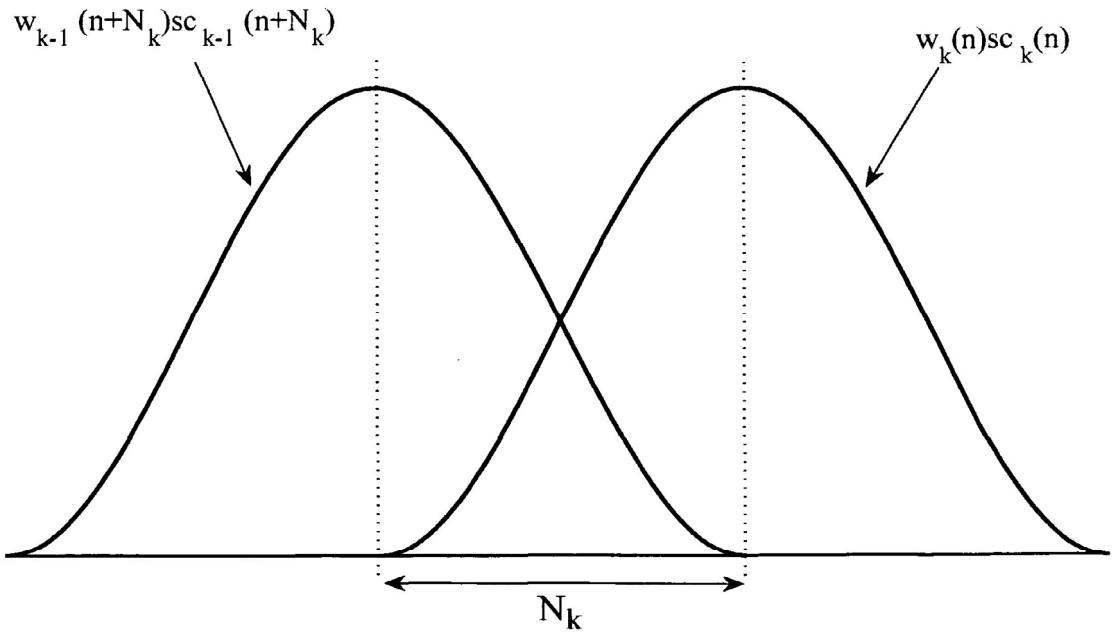

$w_{k-1}(n+N_k)sc_{k-1}(n+N_k)$
$w_k(n)sc_k(n)$
N_k

FIG. 16

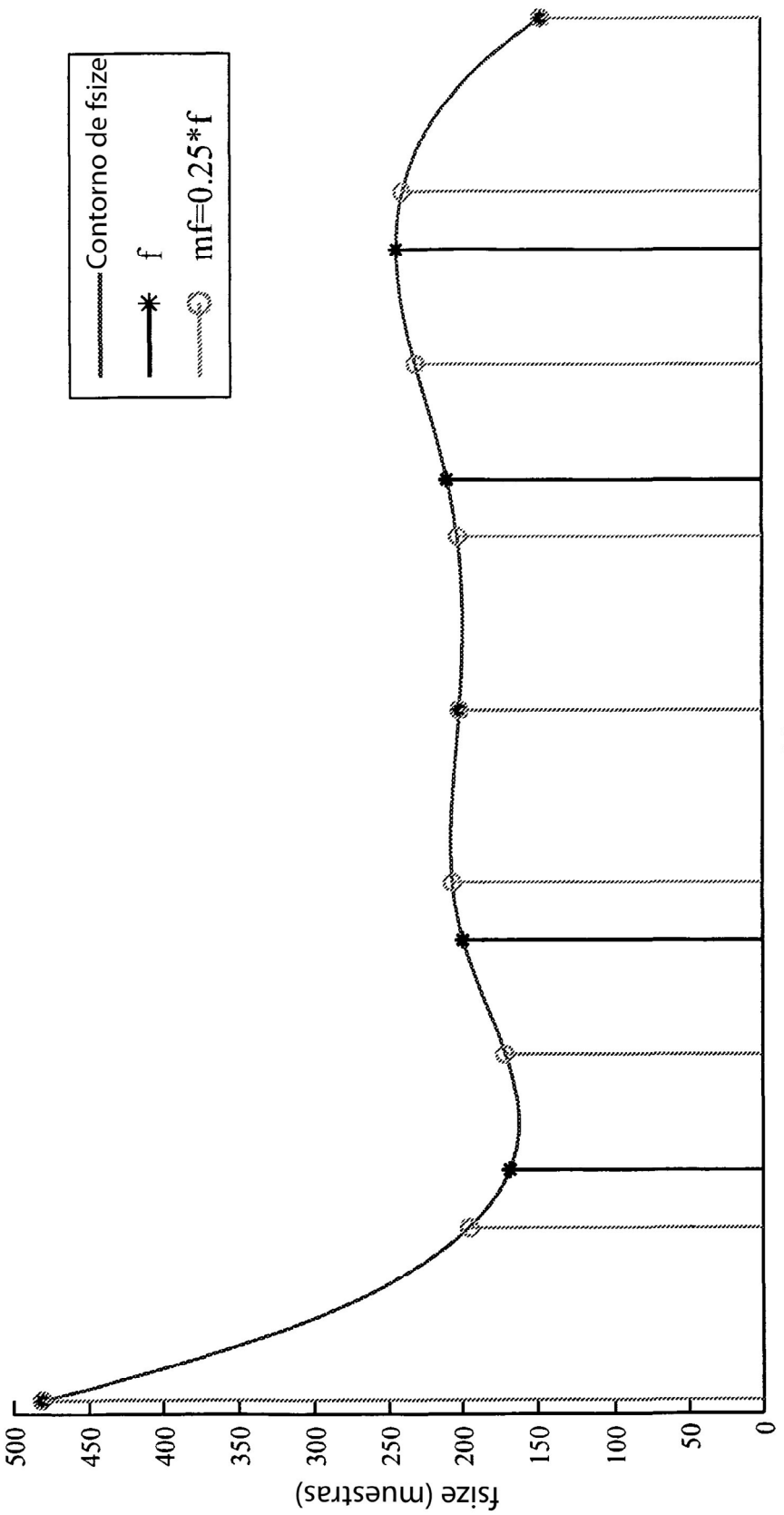

Contorno de fsize
f
mf=0.25*f
fsize (muestras)
500
450
400
350
300
250
200
150
100
50
0
Marcos

FIG. 17

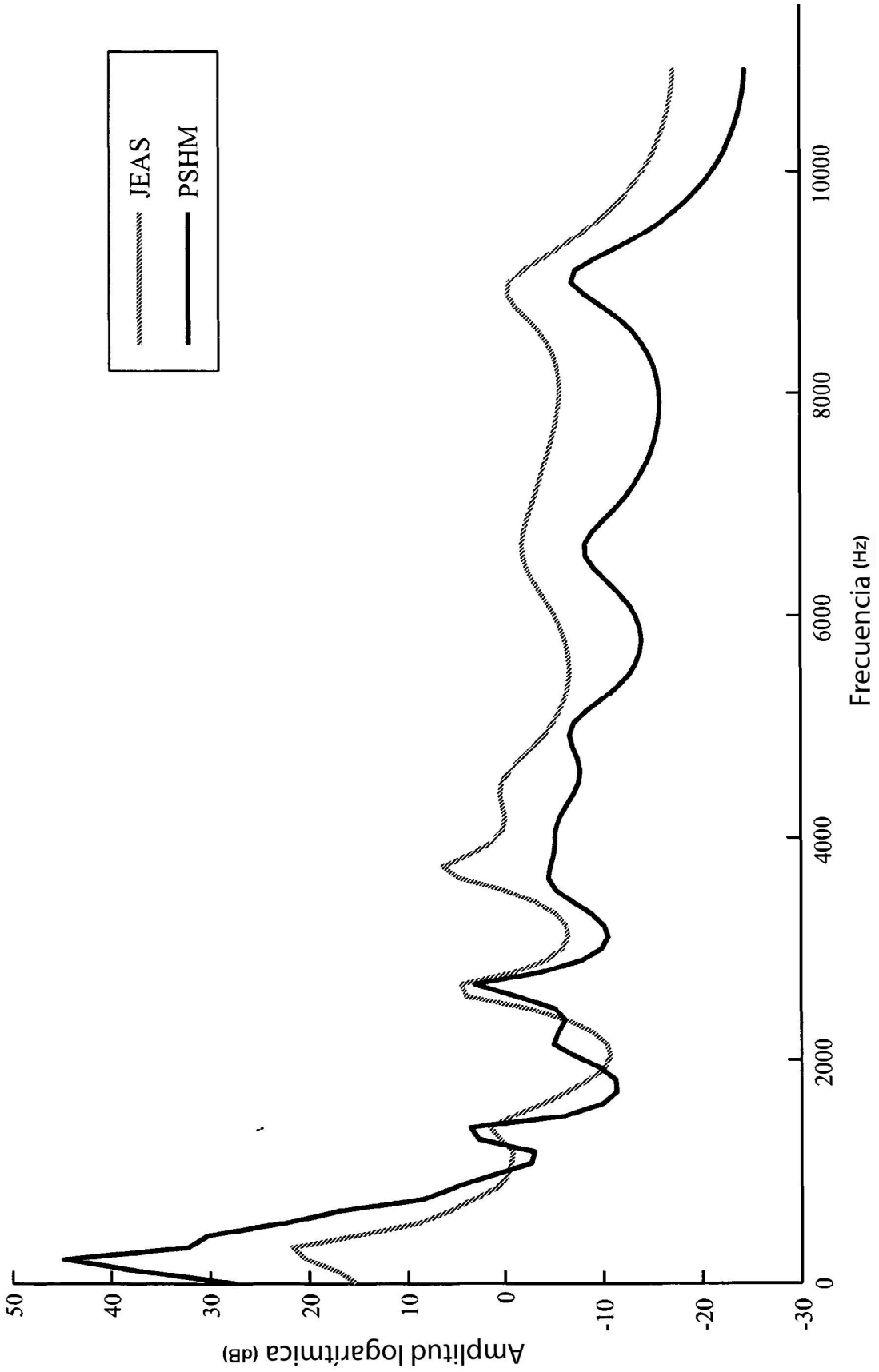

JEAS
PSHM
Amplitud logarítmica (dB)
50
40
30
20
10
0
-10
-20
-30
0
2000
4000
6000
8000
10000
Frecuencia (Hz)

FIG. 18

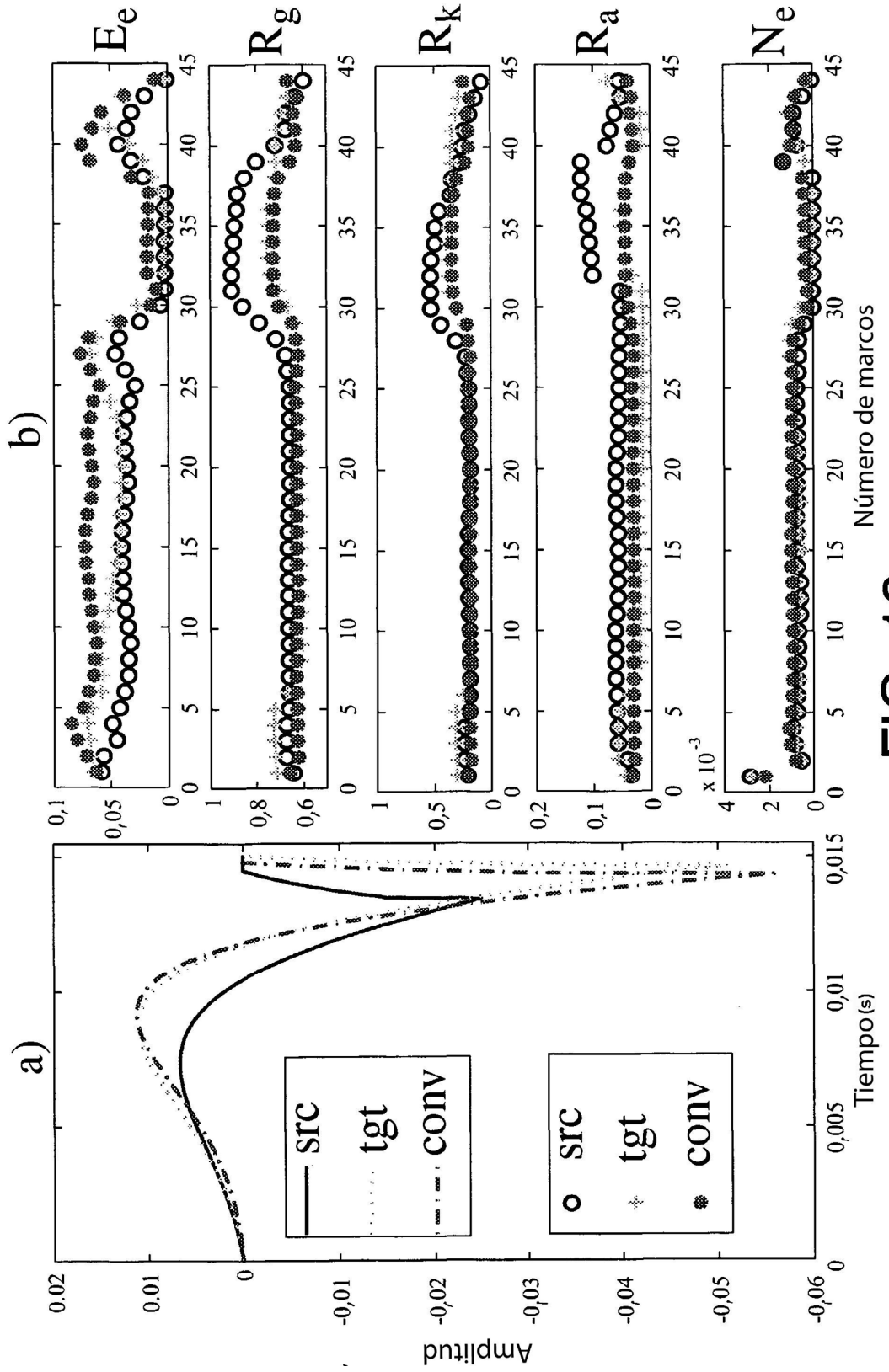

a)
b)
Amplitud
Tiempo (s)
src
tgt
conv
E_e
R_g
R_k
R_a
N_e
Número de marcos
x 10^-3

FIG. 19

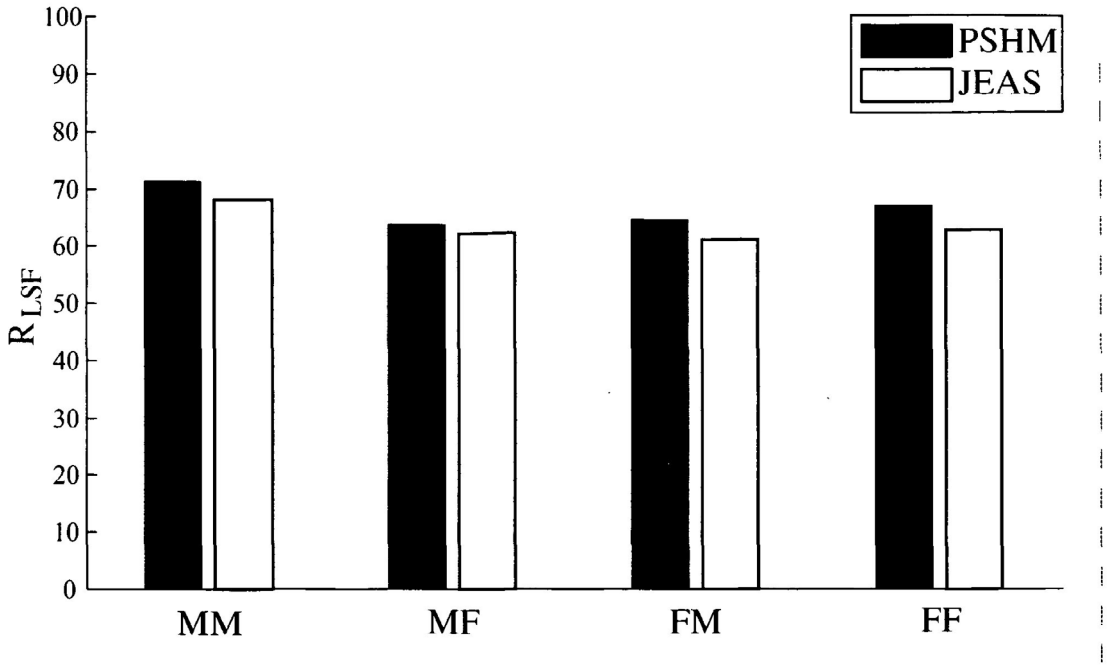

100
90
80
70
60
50
40
30
20
10
0
R_{LSF}
PSHM
JEAS
MM
MF
FM
FF

FIG. 20

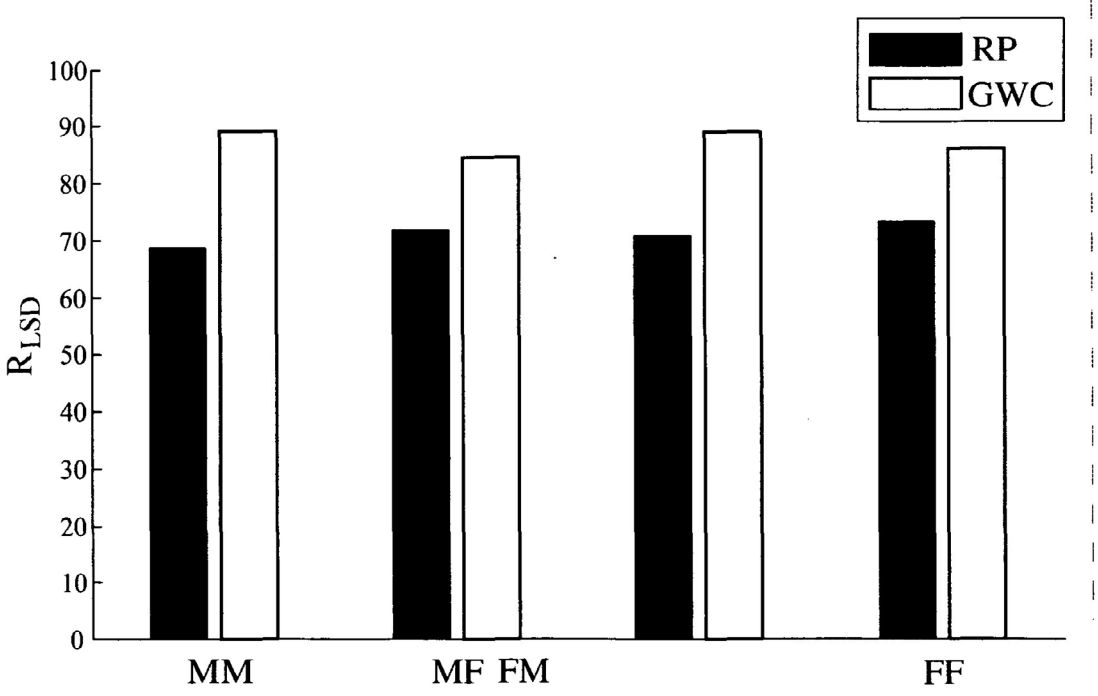

100
90
80
70
60
50
40
30
20
10
0
R_{LSD}
RP
GWC
MM
MF FM
FF

FIG. 21

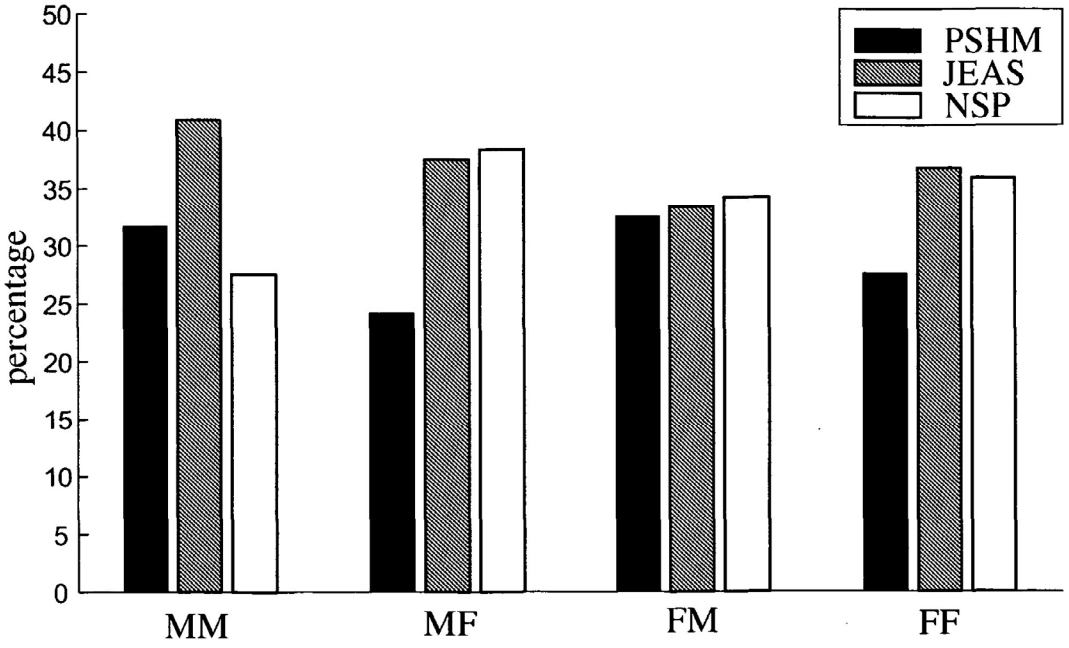

PSHM
JEAS
NSP
percentage
50
45
40
35
30
25
20
15
10
5
0
MM
MF
FM
FF

FIG. 22

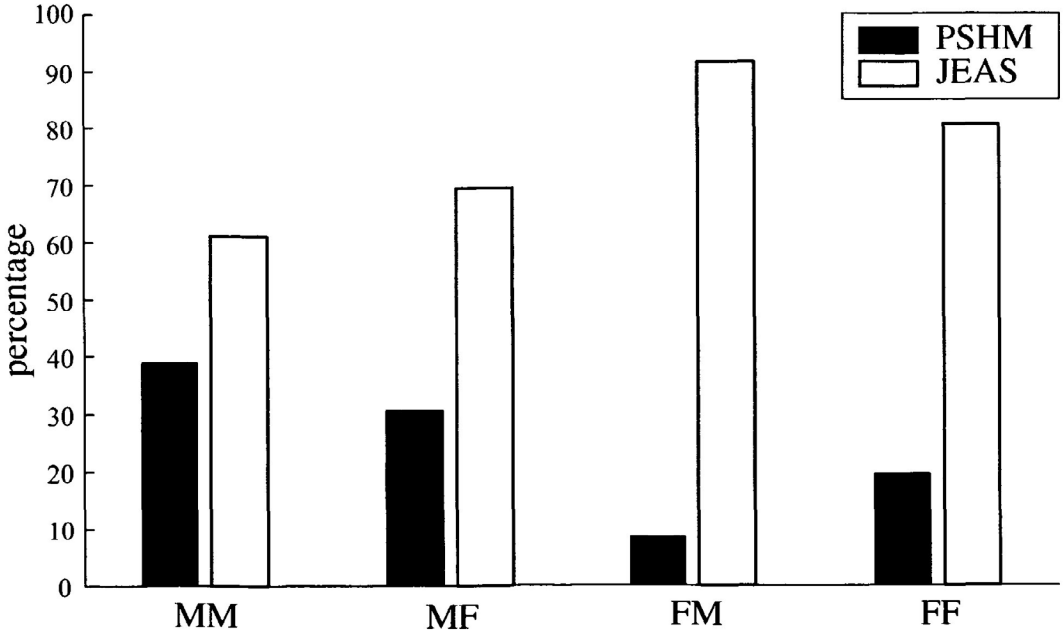

PSHM
JEAS
percentage
100
90
80
70
60
50
40
30
20
10
0
MM
MF
FM
FF

FIG. 23