(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 375 844**

(51) Int. Cl.:
G10L 21/02 (2006.01)
H04R 3/00 (2006.01)
H04B 1/12 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **10167065 .1**
(96) Fecha de presentación: **23.06.2010**
(97) Número de publicación de la solicitud: **2309499**
(97) Fecha de publicación de la solicitud: **13.04.2011**

(54) Título: **PROCEDIMIENTO DE FILTRADO OPTIMIZADO DE LOS RUIDOS NO ESTACIONARIOS CAPTADOS POR UN DISPOSITIVO DE AUDIO MULTIMICRÓFONO, EN PARTICULAR UN DISPOSITIVO TELEFÓNICO "MANOS LIBRES" PARA VEHÍCULO AUTOMÓVIL.**

(30) Prioridad:
22.09.2009 FR 0956506

(45) Fecha de publicación de la mención BOPI:
06.03.2012

(45) Fecha de la publicación del folleto de la patente:
06.03.2012

(73) Titular/es:
Parrot
174 quai de Jemmapes
75010 Paris, FR

(72) Inventor/es:
Vitte, Guillaume;
Seris, Julie y
Pinot, Guillaume

(74) Agente: **Fàbrega Sabaté, Xavier**

ES 2 375 844 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín europeo de patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre concesión de Patentes Europeas).

DESCRIPCIÓN

Procedimiento de filtrado optimizado de los ruidos no estacionarios captados por un dispositivo de audio multimicrófono, en particular un dispositivo telefónico “manos libres” para vehículo automóvil.

La invención se refiere al tratamiento de la voz en un medio ruidoso.

Se refiere en particular, pero de manera no limitativa, al tratamiento de las señales de voz captadas por dispositivos de telefonía para vehículos automóviles.

Estos aparatos incluyen un micrófono (“micro”) sensible que capta no sólo la voz de usuario, sino también el ruido ambiental, ruido que constituye un elemento perturbador que puede ir, en algunos casos, hasta hacer incomprensibles las palabras del hablante. Ocurre lo mismo si se quiere aplicar técnicas de reconocimiento de voz ya que es muy difícil llevar a cabo un reconocimiento de forma en palabras sumergidas en un nivel de ruido elevado.

Esta dificultad ligada a los ruidos ambientales es particularmente molesta en el caso de los dispositivos “manos libres”. En particular, la distancia importante en el micro y el hablante implica un nivel relativo de ruido elevado que dificulta la extracción de la señal útil sumergida en el ruido. Además, el medio muy ruidoso típico del entorno del automóvil presenta características espectrales no estacionarias, es decir que evolucionan de manera imprevisible en función de las condiciones de conducción: paso sobre calzadas deformadas o adoquinadas, autorradio en funcionamiento, etc.

Algunos de estos dispositivos prevén el uso de varios micros, generalmente dos micros, y utilizan la media de las señales captadas, o de otras operaciones más complejas, para obtener una señal con un nivel menor de perturbaciones. En particular, una técnica denominada *beamforming* permite crear por medios de software una directividad que mejora la relación señal/ruido, pero los rendimientos de esta técnica son muy limitados cuando se usan sólo dos micros (concretamente, se estima que tal procedimiento sólo proporciona buenos resultados si se dispone de una red de al menos ocho micros).

Por otra parte, las técnicas clásicas se adaptan sobre todo al filtrado de los ruidos difusos, estacionarios, que proceden de los alrededores del dispositivo y que se encuentran a niveles comparables en las señales captadas por los dos micros.

Por el contrario, un ruido no estacionario o “transitorio”, es decir un ruido que evoluciona de manera imprevisible en función del tiempo, no se discriminará de la voz y, por consiguiente, no se atenuará.

Ahora bien, en el entorno del automóvil, estos ruidos no estacionarios y directivos son muy frecuentes: toque de bocina, paso de una motocicleta, adelantamiento de un coche, etc.

Una dificultad del filtrado de estos ruidos no estacionarios se debe a que sus características temporales y espaciales están muy próximas a las de la voz, de ahí la dificultad por una parte para estimar la presencia de una voz (ya que el hablante no está hablando todo el tiempo) y por otra parte de extraer la señal útil de voz en un entorno muy ruidoso tal como un habitáculo de vehículo automóvil.

Uno de los objetivos de la presente invención es proponer un dispositivo manos libres multimicrófono, en particular, un sistema que aplica sólo dos micrófonos, que permite:

- distinguir de manera eficaz los ruidos no estacionarios de la voz; y
- adaptar la eliminación de ruido a la presencia y a las características de los ruidos no estacionarios detectados, sin modificar la voz eventualmente presente, con el fin de procesar la señal ruidosa de la manera más eficaz.

El punto de partida de la invención consiste en asociar (i) un análisis de coherencia espacial de la señal captada por los dos micros a (ii) un análisis de la dirección de incidencia de estas señales. La invención se basa en efecto en dos constataciones, a saber que:

- la voz presenta generalmente una coherencia espacial superior al ruido; y además que
- la dirección de incidencia de la voz está generalmente bien definida, y se puede suponer que conocida (en el caso de un vehículo automóvil, se define por la posición del conductor, hacia el cual está orientado el micro).

Estas dos propiedades se utilizarán para calcular dos referencias de ruido según procedimientos diferentes:

- una primera referencia de ruido calculada en función de la coherencia espacial de las señales captadas – tal referencia será interesante en la medida en que integra los ruidos no estacionarios poco directivos (fallos en el ronroneo del motor, etc.), y
- una segunda referencia de ruido calculada en función de la dirección principal de incidencia de las señales – esta característica se puede determinar en efecto cuando se usa una red de varios micros (al menos dos), que

conducen a una referencia de ruido que integra sobre todo los ruidos no estacionarios directivos (toques de bocina, paso de una motocicleta, adelantamiento de un coche, etc.).

Estas dos referencias de ruido se utilizarán de manera alterna según la naturaleza del ruido presente, en función de la dirección de incidencia de las señales:

- de manera general, la primera referencia de ruido (la calculada por coherencia espacial) se utilizará por defecto;
- por el contrario, cuando la dirección principal de incidencia de la señal estará alejada de la de la señal útil (la dirección del hablante, supuestamente conocida *a priori*) – es decir en presencia de un ruido directivo bastante potente – la segunda referencia de ruido se utilizará para introducir mayoritariamente en esta última los ruidos no estacionarios directivos y potentes.

Una vez que la referencia de ruido se ha seleccionado de esta manera, se utilizará esta referencia para, por una parte, calcular una probabilidad de ausencia/presencia de voz y, por otra parte, para eliminar el ruido de la señal captada por los micros.

Más concretamente, la invención se refiere, de manera general, a un procedimiento de eliminación de ruido de una señal acústica ruidosa captada por dos micrófonos de un dispositivo de audio mutimicrófono que opera en un medio ruidoso, en particular un dispositivo telefónico “manos libres” para vehículo automóvil. La señal acústica ruidosa comprende una componente útil de voz procedente de una fuente de voz directiva y una componente parásita de ruido, incluyendo esta componente de ruido una componente de ruido lateral no estacionaria directiva.

Tal procedimiento se divulga, por ejemplo, por I. Cohen y B. Berdugo, “*Two-Channel Signal Detection and Speech Enhancement Based on the Transient Beam-to-Reference Ratio*”, Proc. ICASSP 2003, Hong-Kong, páginas 233-236, abril de 2003.

De manera característica de la invención, este procedimiento consta, en el ámbito de frecuencia para una pluralidad de bandas de frecuencias definidas para tramas temporales sucesivas de señal, de las siguientes etapas de procesamiento de señal:

a) cálculo de una primera referencia de ruido por análisis de coherencia espacial de las señales captadas por los dos micrófonos, comprendiendo este cálculo un filtrado lineal predictivo aplicado a las señales captadas por los dos micrófonos y que comprende una sustracción con compensación del desfase entre la señal captada y la señal de salida del filtro predictivo;
b) cálculo de una segunda referencia de ruido por análisis de las direcciones de incidencia de las señales captadas por los dos micrófonos, comprendiendo este cálculo el bloqueo espacial de las componentes de las señales captadas cuya dirección de incidencia se sitúa en el interior de un cono de referencia definido a ambos lados de una dirección predeterminada de incidencia de la señal útil;
c) estimación de una dirección principal de incidencia de las señales captadas por los dos micrófonos;
d) selección como señal de ruido referente de una u otra de las referencias de ruido calculadas en las etapas a) y b) en función de la dirección principal estimada en la etapa c);
e) combinación de las señales captadas por los dos micrófonos en una señal combinada ruidosa;
f) cálculo de una probabilidad de ausencia de voz en la señal combinada ruidosa, a partir de los niveles respectivos de energía espectral de la señal combinada ruidosa y de la señal de ruido referente;
g) a partir de la probabilidad de ausencia de voz calculada en la etapa f) y de la señal combinada ruidosa, reducción selectiva del ruido por aplicación de una ganancia variable propia de cada banda de frecuencias y de cada trama temporal.

Según varias características subsidiarias ventajosas:

- el filtrado lineal predictivo comprende la aplicación de un algoritmo de predicción lineal de tipo mínimos cuadrados medios LMS;
- la estimación de la dirección principal de incidencia de la etapa c) comprende las siguientes subetapas sucesivas: c1) partición del espacio en una pluralidad de sectores angulares; c2) para cada sector, evaluación de un estimador de dirección de incidencia a partir de las señales captadas por los dos micrófonos; y c3) a partir de los valores de estimadores calculados en la etapa c2), estimación de dicha dirección principal de incidencia;
- la selección de la etapa d) es una selección de la segunda referencia de ruido como señal de ruido referente si la dirección principal estimada en la etapa c) se sitúa fuera de un cono de referencia definido a ambos lados de una dirección predeterminada de incidencia de la señal útil;
- la combinación de la etapa e) comprende un prefiltrado de tipo *fixed beamforming*;
- el cálculo de probabilidad de ausencia de voz de la etapa f) comprende la estimación de componentes de ruido pseudoestacionario respectivas contenidas en la señal combinada ruidosa y en la señal de ruido referente, calculándose la probabilidad de ausencia de voz igualmente a partir de estas componentes de ruido pseudoestacionario respectivas;
- la reducción selectiva del ruido de la etapa g) es un procesamiento por aplicación de una ganancia a amplitud espectral logarítmica modificada optimizada OM-LSA.

Se va a describir ahora un ejemplo de aplicación del procedimiento de la invención con referencia a la figura adjunta.

La Figura 1 es un esquema de bloques que muestra los diferentes módulos y funciones aplicadas por el procedimiento de la invención así como sus interacciones.

El procedimiento de la invención se aplica por medios de software, que se pueden descomponer y esquematizar por un cierto número de bloques 10 a 36 ilustrados en la Figura 1.

Estos procesamientos se aplican en forma de algoritmos apropiados ejecutados por un microcontrolador o un procesador digital de señal. Aunque, por motivos de claridad, estos diversos procesamientos se presentan en forma de distintos módulos, aplican elementos comunes y corresponden en la práctica a una pluralidad de funciones globalmente ejecutadas por un mismo software.

La señal de la que se desea eliminar el ruido procede de una pluralidad de señales captadas por una red de micros (que, en la configuración mínima, puede ser simplemente una red de dos micros, como en el ejemplo ilustrado) dispuestos según una configuración predeterminada. En la práctica, estos dos micros se pueden, por ejemplo, instalar en el techo interior de un habitáculo de coche, a aproximadamente 5 cm el uno del otro; y tener el lóbulo principal de su diagrama de directividad orientado hacia el conductor. Esta dirección, considerada como *a priori* conocida, se designará dirección de incidencia de la señal útil.

Se denominará “ruido lateral” un ruido no estacionario directivo cuya dirección de incidencia está alejada de la de la señal útil, y se denominará “cono privilegiado” la dirección o sector angular del espacio donde se encuentra la fuente de señal útil (la voz del hablante) respecto de la red de micros. Cuando una fuente sonora se manifieste fuera del cono privilegiado, se tratará por lo tanto de un ruido lateral, que se deseara atenuar. Como se ilustra en la Figura 1, las señales ruidosas captadas por los dos micros $x_1(n)$ y $x_2(n)$ son objeto de una transposición en el domino de frecuencia (bloques 10) por un cálculo de transformada de Fourrier a corto plazo (FFT) cuyo resultado se denota respectivamente como $X_1(k,l)$ y $X_2(k,l)$, siendo k el índice de la banda de frecuencia y siendo l el índice de la trama temporal. Las señales procedentes de los dos micros se aplican asimismo a un módulo 12 que lleva a cabo un algoritmo LMS predictivo esquematizado por el bloque 14 y que proporciona, después del cálculo de una transformada de Fourrier a corto plazo (bloque 16) una señal $Y(k, l)$ que servirá para el cálculo de una primera referencia de ruido $Ref_1(k,l)$ ejecutado por un bloque 18, esencialmente con un criterio de coherencia espacial.

Otra referencia de ruido $Ref_2(k,l)$ se calcula mediante un bloque 20, esencialmente con un criterio de bloqueo angular), a partir de las señales $X_1(k,l)$ y $X_2(k,l)$ directamente obtenidas, en el ámbito de frecuencia, a partir de las señales $x_1(n)$ y $x_2(n)$.

Un bloque 22 lleva a cabo la selección de una u otra de las referencias de ruido $Ref_1(k,l)$ o $Ref_2(k,l)$ en función del resultado de un cálculo del ángulo de incidencia de las señales llevado a cabo por el bloque 24 a partir de las señales $X_1(k,l)$ y $X_2(k,l)$. La referencia de ruido elegida, $Ref(k,l)$, se utiliza como canal de ruido referente de un bloque 26 de cálculo de una probabilidad de ausencia de voz llevada a cabo en una señal ruidosa $X(k,l)$ que resulta de una combinación, llevada a cabo por el bloque 28, de las dos señales $X_1(k,l)$ y $X_2(k,l)$. El bloque 26 tiene asimismo en cuenta las componentes de ruido pseudoestacionario respectivas del canal de ruido referente y de la señal ruidosa, componentes estimadas por los bloques 30 y 32.

El resultado $q(k,l)$ del cálculo de probabilidad de ausencia de voz y la señal ruidosa $X(k,l)$ se aplican en la entrada de un algoritmo de control de ganancia OM-LSA (bloque 34) cuyo resultado l $\hat{S}(k,l)$ se somete (bloque 36) a una transformación de Fourrier inversa (iFFT) para obtener en el ámbito temporal una estimada $\hat{S}(t)$ de la señal de voz sin ruido.

Ahora se describirá en detalle cada una de las etapas del procesamiento.

Transformada de Fourier de las señales captadas por los micros (bloques 10)

La señal en el ámbito temporal $x_n(t)$ procedente de cada uno de los N micros (N = 1,2 en el ejemplo ilustrado) se digitaliza, se divide en tramas de T puntos temporales, ventana temporalmente por una ventana de tipo Hanning y, a continuación, la transformada de Fourier rápida FFT (transformada a corto plazo) $X_n(k,l)$ se calcula para cada una de estas señales:

$$X_n(k,l) = a_n.d_n(k) \times S(k,l) + V_n(k,l)$$

con

$$d_n(k) = e^{-i2\pi f_k \tau_n}$$

l siendo el índice de la trama temporal,

k	siendo el índice de la banda de frecuencias, y
f_k	siendo la frecuencia central de la banda de frecuencias de índice por k
$S(k,l)$	designando la fuente de señal útil,
a_n y τ_n	designando la atenuación y el retardo experimentados por la señal útil captado al nivel del micro n, y
$V_n(k,l)$	designando el ruido captado por el micro n.

Cálculo de una primera referencia de ruido por coherencia espacial (bloque 12)

La idea fundamental sobre la cual se basa la invención es que, en un entorno de telecomunicaciones, la voz es una señal emitida por una fuente bien localizada, relativamente próxima a los micros y casi enteramente captada en camino directo. Por el contrario, los ruidos estacionarios y no estacionarios, que proceden sobre todo de los alrededores del usuario, se pueden asociar a fuentes alejadas, en gran número y con una correlación estadística inferior a la voz entre los dos micros.

En un entorno de telecomunicaciones, la voz es por lo tanto más coherente espacialmente que el ruido.

Partiendo de este principio, es posible explotar la propiedad de coherencia espacial para construir un canal de ruido de referencia más rico y más adaptado que con un *beamformer*. El sistema prevé con este fin la utilización de un filtro predictivo 14 de tipo LMS (*Least Mean Squares*, mínimos cuadrados medios) que tiene por entradas las señales $x_1(n)$ y $x_2(n)$ captadas por el par de micros. Se denotará como $y(n)$ la salida del LMS y $e(n)$ el error de predicción.

Este filtro predictivo se utiliza para predecir a partir de $x_2(n)$ la componente de voz que se encuentra en $x_1(n)$. En efecto, al ser espacialmente más coherente, la voz se predecirá mejor por el filtro adaptativo que el ruido.

Una primera posibilidad consiste en tomar para el canal de ruido referente la transformada de Fourier del error de predicción:

$$E(k,l) = X_1(k,l) - Y(k,l)$$

siendo $E(k,l)$, $X_1(k,l)$ e $Y(k,l)$ las transformadas de Fourier a corto plazo (TFCT) respectivas de $e(k,l)$, $x_1(k,l)$ e $y(k,l)$.

Sin embargo, en la práctica se constata un cierto desfase entre $X_1(k,l)$ e $Y(k,l)$ debido a una convergencia imperfecta del algoritmo LMS, lo que impide una buena discriminación ente voz y ruido.

Para paliar este defecto, es posible definir la primera señal de ruido referente $Ref_1(k,l)$ como:

$$Ref_1(k,l) = X_1(k,l) - X_1(k,l)\frac{|Y(k,l)|}{|X_1(k,l)|}$$

A diferencia de numerosos procedimientos clásicos de estimación del ruido, no se utiliza ninguna hipótesis de estacionaridad sobre el ruido para calcular este primer canal de referencia $Ref_1(k,l)$. Una de las ventajas es, por consiguiente, que este canal de ruido integre una parte de los ruidos no estacionarios, en particular los que tienen una baja correlación estadística y que no son predecibles entre los dos micros.

Cálculo de una segunda referencia de ruido por bloqueo espacial (bloque 20)

En un entorno de telecomunicaciones, es posible encontrar ruidos cuya fuente esté bien localizada y relativamente próxima a los micros. Se trata en general de ruidos puntuales bastante potentes (paso de una motocicleta, adelantamiento de un coche, etc.) y que pueden ser molestos.

Las hipótesis utilizadas para el cálculo del primer canal de ruido referente no se verifican en este tipo de ruido; por el contrario, estos ruidos tienen la particularidad de tener una dirección de incidencia bien definida y distinta de la dirección de incidencia de la voz.

Para utilizar esta propiedad, se supondrá que el ángulo de incidencia θ_S de la voz es conocido, por ejemplo definido como el ángulo entre la mediatriz del par de micros y la dirección de referencia que corresponde a la fuente de voz útil.

Más concretamente, se lleva a cabo una partición del espacio en sectores angulares que describen el espacio, y de los que cada uno corresponde a una dirección definida por un ángulo θ_j, j $\in$ [1,M] on por ejemplo M = 19, proporcionando la colección de ángulos {-90º, -80º …, 0º, ... +80º, +90º}. Se observará que no hay ninguna conexión entre el número N de micros y el número M de ángulos probados: por ejemplo, es totalmente posible probar M = 19 ángulos con un solo par de micros (N=2).

Se proporciona la partición *{A,I}* de los ángulos θ_j que están respectivamente "autorizados" y "prohibidos", estando los ángulos $\theta_a \in A$ "autorizados" porque corresponden a señales procedentes de un cono privilegiado centrado en θ_S, mientras que los ángulos $\theta_i \in I$ están "prohibidos" porque corresponden a ruidos laterales indeseables.

El segundo canal de ruido referente $Ref_2(k,l)$ se define de la siguiente manera:

$$Ref_2(k,l) = \frac{1}{|A|}\sum_{\theta_a \in A}\left(X_1(k,l) - X_2(k,l) \times e^{\frac{i2\pi . f_k . d . \sin\theta_a}{c}} \right)$$

$X_1(k,l)$ siendo la TFCT de la señal registrada por el micro de índice 1,
$X_2(k,l)$ siendo la TFCT de la señal registrada por el micro de índice 2,
f_k siendo la frecuencia central de la banda de frecuencias k,
l siendo la trama,
d siendo la distancia entre los dos micros,
c siendo la celeridad del sonido, y
$|A|$ siendo el número de ángulos "autorizados" del cono privilegiado.

En cada término de esta suma, se resta a la señal del micro de índice 1 la señal del micro de índice 2 desfasada en un ángulo θ_a que pertenece a A (subcolección de los ángulos "autorizados"). De este modo, en cada término se bloquean espacialmente las señales que tienen una dirección de propagación θ_a "autorizada". Este bloqueo espacial se lleva a cabo para todos los ángulos autorizados.

En este segundo canal de ruido referente $Ref_2(k,l)$, se deja por lo tanto pasar los eventuales ruidos laterales (ruidos no estacionarios directivos) bloqueando espacialmente la señal de voz.

Elección de la referencia de ruido en función de la dirección de incidencia de las señales (bloques 22 y 24)

Esta selección implica una estimación del ángulo de incidencia $\hat{\theta}(k,l)$ de las señales.

$$\hat{\theta}(k,l) = \underset{\theta_j, j\in[1,M]}{\arg\max}\left\| P_{1,2}(\theta_j,k,l) \right\|$$

con

$$P_{1,2}(\theta_j,k,l) = E\left(X_1(k,l).\overline{X}_2(k,l).e^{-i2\pi f_k \tau_j} \right)$$

y

$$\tau_j = \frac{d}{c}\operatorname{sen}\theta_j$$

El canal de ruido referente seleccionado *Ref(k,l)* va a depender de la detección de un ángulo "autorizado" o "prohibido" para la trama l y la banda de frecuencia k:

- si $\theta(k,l)$ está "autorizado" ($\theta(k,l) \in A$), entonces $Ref(k,l) = Ref_1(k,l)$
- *si* $\theta(k,l)$ está "prohibido" ($\theta(k,l) \in I$), entonces $Ref(k,l) = Ref_2(k,l)$
- *si* $\theta(k,l)$ no está definido, entonces $Ref(k,l) = Ref_1(k,l)$.

Así, en el caso de un ángulo "autorizado" detectado, o en ausencia de señales directivas en la entrada de los micros, el canal de ruido referente *Ref(k,l)* se calcula por coherencia espacial, lo que permite integrar los ruidos no estacionarios poco directivos.

Por el contrario, si se detecta un ángulo "prohibido", esto significa que un ruido directivo y bastante potente está presente. En este caso, el canal de ruido referente *Ref(k,l)* se calcula según un procedimiento diferente, por bloqueo espacial, de manera que se introducen eficazmente en este canal los ruidos no estacionarios directivos y potentes.

Constitución de una señal combinada con ruido parcialmente eliminado (bloque 28)

Las señales $X_n(k,l)$ (las TFCT de las señales captadas por los micros) se pueden combinar entre sí mediante una técnica sencilla de prefiltrado por *beamforming* del tipo *Delay and Sum*, que se aplica para obtener una señal combinada $X(k,l)$ con ruido parcialmente eliminado:

$$X(k,l) = \frac{1}{2}\left[X_1(k,l) + \overline{d_2(k)}.X_2(k,l)\right]$$

con

$$d_2(k) = e^{i2\pi f_k \tau_s} \quad \text{con} \quad \tau_s = \frac{d}{c}\,\text{sen}\,\theta_s$$

Cuando el sistema considerado consta, como en el presente ejemplo, de dos micros cuya mediatriz corta la fuente, el ángulo θ_S es nulo y se trata de una simple media que se realiza sobre los dos micros. Por otra parte hay que resaltar que, concretamente, al ser el número de micros limitado, este procesamiento sólo proporciona una pequeña mejora de la relación señal/ruido, del orden de sólo 1 dB.

Estimación del ruido pseudoestacionario (bloques 30 y 32)

Esta etapa tiene por objeto calcular una estimación de la componente de ruido pseudoestacionario presente en la referencia de ruido *Ref(k,l)* (bloque 30) y, de la misma manera, la componente de ruido pseudoestacionario presente en la señal en la que hay eliminar el ruido *X(k,l)* (bloque 32). Existe un gran número de publicaciones acerca de este tema, siendo en efecto la estimación de la componente de ruido pseudoestacionario un problema clásico bastante bien resuelto. Diferentes procedimientos son eficaces y se pueden utilizar con este fin, en particular un algoritmo de estimación de la energía de la componente de ruido pseudoestacionario con promediación recursiva mediante el control de los mínimos (MCRA) como el descrito por I. Cohen y B. Berdugo, "*Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement*", *IEEE Signal Processing Letters*, Vol. 9, nº 1, páginas 12-15, enero de 2002.

Cálculo de la probabilidad de ausencia de voz (bloque 26)

Un procedimiento eficaz y reconocido para estimar la probabilidad de ausencia de voz en un entorno ruidoso es el de la relación de los transitorios, descrita por I. Cohen y B. Berdugo, "*Two-Channel Signal Detection and Speech Enhancement Based on the Transient Beam-to-Reference Ratio*", Proc. ICASSP 2003, Hong-Kong, páginas 233-236, abril de 2003. La relación de los transitorios se define de la siguiente manera:

$$\Omega(k,l) = \frac{S[X(k,l)] - M[X(k,l)]}{S[Ref(k,l)] - M[Ref(k,l)]}$$

X(k,l) siendo la señal combinada con ruido parcialmente eliminado,
Ref(k,l) siendo el canal de ruido referente calculado en la parte precedente,
k siendo la banda de frecuencias, y
l siendo la trama,

El operador *S* es una estimación de la energía instantánea, y el operador *M* es una estimación de la energía pseudoestacionaria (estimación efectuada por los bloques 30 y 32). *S – M* proporciona una estimación de las partes transitorias de la señal analizada, denominadas también transitorios.

Las dos señales analizadas son aquí la señal ruidosa combinada *X(k,l)* y la señal del canal de ruido referente *Ref(k,l)*. El numerador pondrá entonces de manifiesto los transitorios de voz y de ruidos, mientras que el denominador va a extraer solamente los transitorios de ruidos que se encuentran en el canal de ruido referente.

Así, en presencia de voz pero en ausencia de ruido no estacionario, la relación $\Omega(k,l)$ va a tender hacia un límite alto $Q_{max}(k)$ mientras que, por el contrario, en ausencia de voz pero en presencia de ruido no estacionario, esta relación se a aproximar al límite bajo $\Omega_{min}(k)$, siendo *k* la banda de frecuencias. Esto va a permitir realizar la diferenciación entre voz y ruidos no estacionarios.

En el caso general, tenemos:

$$\Omega_{\min}(k) \le \Omega(k,l) \le \Omega_{\max}(k)$$

La probabilidad de ausencia de voz, denotada aquí *q(k,l)* se va a calcular de la siguiente manera.

Para cada trama l y cada banda de frecuencias k:

i) *Cálculo de* $S[X(k,l)]$, $S[Ref(k,l)]$, $M[X(k,l)]$ *y* $M[Ref(k,l)]$;

ii) *Si* $S[X(k,l)] \ge \alpha_X M[X(k,l)]$, *la voz es susceptible de estar presente, el análisis continúa en la etapa (iii), en*

caso contrario, la voz está ausente: entonces q(k,l) = 1.

iii) *Si S[Ref(k,l)] ≥ α_{Ref}M[Ref(k,l)], el ruido transitorio es susceptible de estar presente, el análisis continúa en la etapa (iv), en caso contrario, esto significa que los transitorios encontrados en X(k,l) son todos transitorios de voz: entonces q(k,l) = 0;*

iv) *Cálculo de la relación* $\Omega(k,l) = \dfrac{S[X(k,l)] - M[X(k,l)]}{S[Ref(k,l)] - M[Ref(k,l)]}$;

v) *Determinación de la probabilidad de ausencia de voz:*

$$q(k,l) = \max\left(\min\left(\frac{\Omega_{\max}(k,l) - \Omega(k,l)}{\Omega_{\max}(k,l) - \Omega_{\min}(k,l)}, 1\right), 0\right) .$$

Las constantes α_X y α_{Ref} utilizadas en este algoritmo son de hecho umbrales de detección de las partes transitorias. Los parámetros α_X, α_{Ref} así como $\Omega_{min}(k)$ y $\Omega_{max}(k)$, se eligen todos de manera que se correspondan con situaciones típicas, próximas a la realidad.

Reducción de ruido por aplicación de una ganancia OM-LSA (bloque 34)

La probabilidad *q(k,l)* de ausencia de voz calculada en el bloque 26 se va a utilizar como parámetro de entrada en una técnica (en si misma conocida) de eliminación de ruido. Presenta la ventaja de permitir la identificación de los periodos de ausencia de voz incluso en presencia de un ruido no estacionario, poco directivo o directivo. La probabilidad de ausencia de voz es un estimador crucial para el buen funcionamiento de una estructura de eliminación de ruido tal como la que se va a utilizar, ya que subtiende la buena estimación del ruido y el cálculo de una ganancia de eliminación de ruido eficaz.

Se puede utilizar ventajosamente un procedimiento de eliminación de ruido del tipo OM-LSA (*Optimally Modified – Log Spectral Amplitude*) tal como el descrito por: I. Cohen "*Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator*", *IEEE Signal Processing Letters*, Vol. 9, nº 4, abril de 2002.

Esencialmente, la aplicación de una ganancia denominada "ganancia LSA" (*Log-Spectral Amplitude*) permite minimizar la distancia cuadrática media entre el logaritmo de la amplitud de la señal estimada y el logaritmo de la amplitud de la señal de voz original. Este segundo criterio se muestra superior al primero ya que la distancia elegida es más adecuada para el comportamiento del oído humano y proporciona por lo tanto cualitativamente mejores resultados. En todos los casos, la idea esencial es disminuir la energía de las componentes de frecuencia muy parasitadas aplicándoles una ganancia reducida, dejando a la vez intactas (por aplicación de una ganancia igual a 1) las que lo están poco o nada.

El algoritmo "OM-LSA" (*Optimally-Modified Log-Spectral Amplitude)* mejora el cálculo de la ganancia LSA que se va a aplicar, ponderándolo por la probabilidad condicional de presencia de voz.

En este procedimiento, la probabilidad de ausencia de voz interviene en dos momentos importantes, para la estimación de la energía del ruido y para el cálculo de la ganancia final, y la probabilidad *q(k,l)* se utilizará a estos dos niveles. Si se denota $\lambda_{Ruido}(k,1)$ la estimación de la densidad espectral de potencia del ruido, esta estimación viene dada por:

$$\hat{\lambda}_{Ruido}(k,l) = \alpha_{Ruido}(k,l).\hat{\lambda}_{Ruido}(k,l-1) + \left[1 - \alpha_{Ruido}(k,l)\right]\left|X(k,l\right|^2$$

con

$$\alpha_{Ruido}(k,l) = \alpha_B + (1-\alpha_B).p_{spa}(k,l)$$

Se puede observar aquí que la probabilidad *q(k,l)* modula el factor de olvido en la estimación del ruido, que se actualiza más rápidamente en la señal ruidosa *X(k,l)* cuando la probabilidad de ausencia de voz es alta, condicionando este mecanismo totalmente la calidad de $\lambda_{Ruido}(k,l)$.
La ganancia de eliminación de ruido $G_{OM\text{-}LSA}(k,l)$ viene dada por:

$$G_{OM-LSA}(k,l) = \left\{G_{H1}(k,l)\right\}^{1-q(k,l)}.G_{\min}^{q(k,l)}$$

$G_{H1}(k,l)$ siendo una ganancia de eliminación de ruido (cuyo cálculo depende de la estimación del ruido λ_{Ruido}) descrita en el artículo de Cohen anteriormente mencionado, y

G_{min} siendo una constante correspondiente a la eliminación de ruido aplicada cuando se considera la voz como ausente.

Se observará que la probabilidad *q(k,l)* desempeña aquí una función importante en la determinación de la ganancia $G_{OM\text{-}LSA}(k,l)$. En particular, cuando esta probabilidad es nula, la ganancia es igual a G_{min} y se aplica una reducción de ruido máxima: si, por ejemplo, se elige un valor de 20 dB para G_{min}, los ruidos no estacionarios detectados anteriormente se atenúan en 20 dB.

La señal sin ruidos $\hat{S}(k,l)$ a la salida del bloque 34 viene dada por:

$$\hat{S}(k,l) = G_{OM-LSA}(k,l).X(k,l)$$

Se observará que comúnmente tal estructura de eliminación de ruido produce un resultado poco natural y agresivo sobre los ruidos no estacionarios, que se confunden con la voz útil. Uno de los mayores intereses de la invención es, por el contrario, eliminar eficazmente estos ruidos no estacionarios.

Por otra parte, en una variante ventajosa, es posible utilizar en las expresiones dadas más arriba, una probabilidad de ausencia de voz híbrida $q_{híbrida}(k,l)$ que se calculará con la ayuda de *q(k,l)* y de otra probabilidad de ausencia de voz $q_{std}(k,l)$, por ejemplo evaluada según el procedimiento descrito en el documento WO 2007/099222 A1 (Parrot SA).

Tenemos entonces:

$$q_{hibrida}(k,l) = \max\left(q(k,l),\ q_{std}(k,l)\right)$$

Reconstitución temporal de la señal (bloque 36)

La última etapa consiste en aplicar a la señal $\hat{S}(k,l)$ una transformada de Fourier rápida inversa iFTT para obtener en el ámbito temporal la señal de voz sin ruido $\hat{S}(t)$ buscada.

REIVINDICACIONES

1. Procedimiento de eliminación de ruido de una señal acústica ruidosa captada por dos micrófonos de un dispositivo de audio multimicrófono que opera en un medio ruidoso, en particular un dispositivo telefónico "manos libres" para vehículo automóvil,
comprendiendo la señal acústica ruidosa un componente útil de voz procedente de una fuente de voz directiva y un componente parásito de ruido, incluyendo este componente de ruido un componente de ruido lateral no estacionario direccional,
procedimiento **caracterizado por que** comprende, en el ámbito de frecuencia para una pluralidad de bandas de frecuencias definidas para tramas temporales sucesivas de señal, las siguientes etapas de procesamiento de la señal:

a) cálculo (18) de una primera referencia de ruido por análisis de coherencia espacial de las señales captadas por los dos micrófonos, comprendiendo este cálculo un filtrado lineal predictivo aplicado a las señales captadas por los dos micrófonos y que comprende una sustracción con compensación del desfase entre la señal captada y la señal de salida del filtro predictivo;
b) cálculo (20) de una segunda referencia de ruido por análisis de las direcciones de incidencia de las señales captadas por los dos micrófonos, comprendiendo este cálculo el bloqueo espacial de las componentes de las señales captadas cuya dirección de incidencia se sitúa en el interior de un cono de referencia definido a ambos lados de una dirección predeterminada de incidencia de la señal útil;
c) estimación (24) de una dirección principal de incidencia ($\hat{\theta}(k,l)$) de las señales captadas por los dos micrófonos;
d) selección (22) como señal de ruido referente ($Ref(k,l)$) de una u otra de las referencias de ruido calculadas en las etapas a) y b), en función de la dirección principal estimada en la etapa c);
e) combinación (28) de las señales captadas por los dos micrófonos en una señal combinada ruidosa ($X(k,l)$);
f) cálculo (26) de una probabilidad de ausencia de voz ($q(k,l)$) en la señal combinada ruidosa, a partir de niveles respectivos de energía espectral de la señal combinada ruidosa ($X(k,l)$) y de la señal de ruido referente ($Ref(k,l)$);
g) a partir de la probabilidad de ausencia de voz ($q(k,l)$) calculada en la etapa f) y de la señal combinada ruidosa ($X(k,l)$), reducción selectiva del ruido (34) mediante la aplicación de una ganancia variable propia de cada banda de frecuencias y de cada trama temporal.

2. El procedimiento de la reivindicación 1, en el que el filtrado predictivo comprende la aplicación de un algoritmo de predicción lineal de tipo mínimos cuadrados medios LMS.

3. El procedimiento de la reivindicación 1, en el que la estimación (24) de la dirección principal de incidencia de la etapa c) comprende las siguientes subetapas sucesivas:

c1) partición del espacio en una pluralidad de sectores angulares;
c2) para cada sector, evaluación de un estimador de dirección de incidencia a partir de las dos señales captadas por los dos micrófonos correspondientes; y
c3) a partir de los valores de estimadores calculados en la etapa c2), estimación de dicha dirección principal de incidencia.

4. El procedimiento de la reivindicación 1, en el que la selección (22) de la etapa d) es una selección de la segunda referencia de ruido como señal de ruido referente si la dirección principal estimada en la etapa c) se sitúa fuera de un cono de referencia definido a ambos lados de una dirección predeterminada de incidencia de la señal útil.

5. El procedimiento de la reivindicación 1, en el que la combinación (28) de la etapa e) comprende un prefiltrado de tipo *fixed beamforming.*

6. El procedimiento de la reivindicación 1, en el que el cálculo (26) de probabilidad de ausencia de voz de la etapa f) comprende la estimación (30, 32) de componentes de ruido pseudoestacionario respectivos contenidos en la señal combinada ruidosa y en la señal de ruido referente, calculándose asimismo la probabilidad de ausencia de voz ($q(k,l)$) a partir de estos componentes de ruido pseudoestacionario respectivos.

7. El procedimiento de la reivindicación 1, en el que la reducción selectiva del ruido (34) de la etapa g) es un procesamiento mediante aplicación de una ganancia de amplitud espectral logarítmica modificada optimizada OM-LSA.

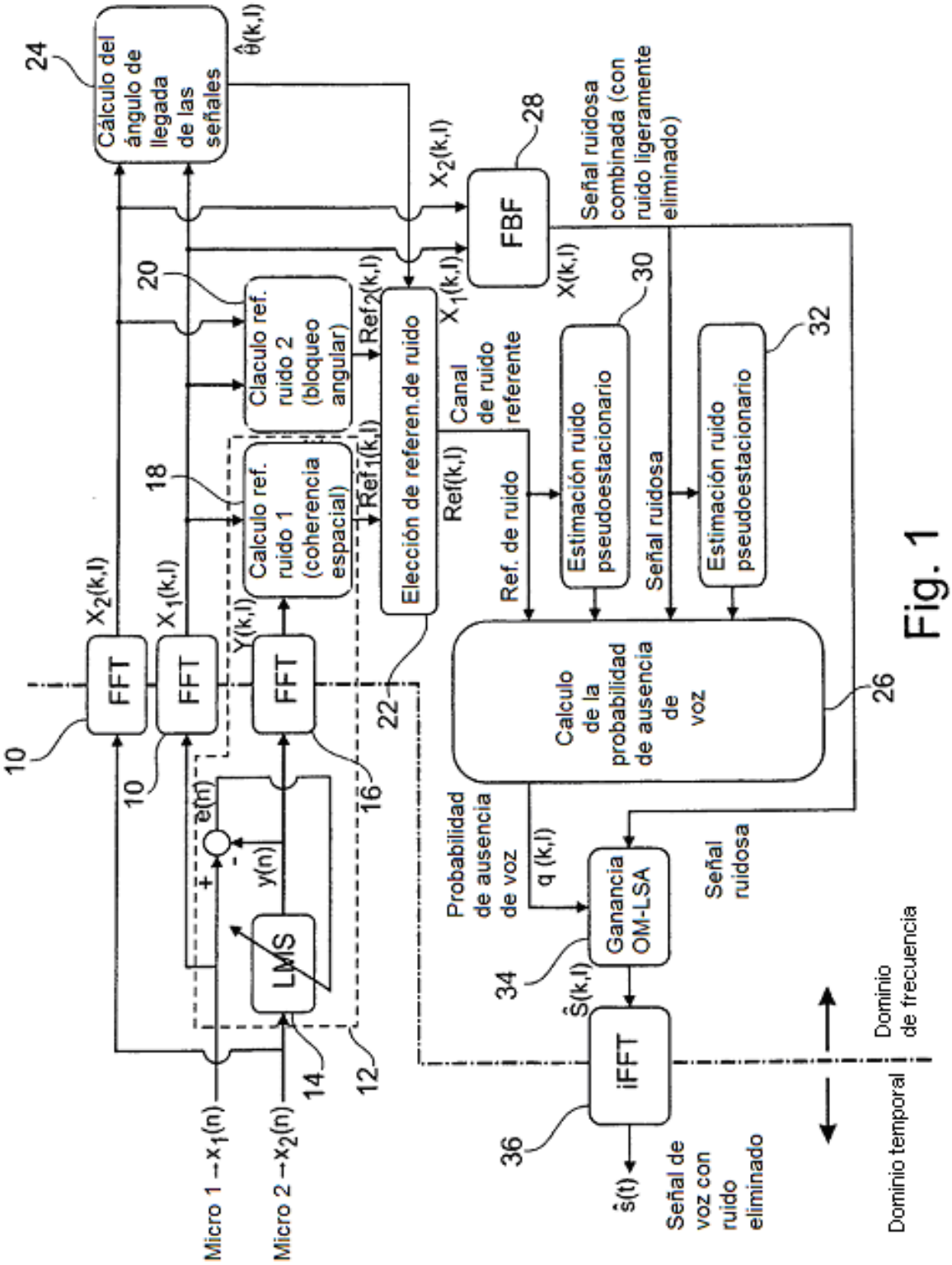
Cálculo del ángulo de llegada de las señales
$\hat{\theta}(k,l)$
24
FBF
28
$X_2(k,l)$
$X_1(k,l)$
$X(k,l)$
Señal ruidosa combinada (con ruido ligeramente eliminado)
20
Claculo ref. ruido 2 (bloqueo angular)
$Ref_2(k,l)$
Elección de referen.de ruido
Canal de ruido referente
Ref(k,l)
Ref. de ruido
30
Estimación ruido pseudoestacionario
Señal ruidosa
32
Estimación ruido pseudoestacionario
18
Calculo ref. ruido 1 (coherencia espacial)
$Ref_1(k,l)$
$Y(k,l)$
22
Calculo de la probabilidad de ausencia de voz
26
FFT
10
FFT
10
FFT
16
e(n)
y(n)
LMS
14
12
Micro 1 → $x_1(n)$
Micro 2 → $x_2(n)$
Probabilidad de ausencia de voz
q (k,l)
Ganancia OM-LSA
34
Señal ruidosa
$\hat{S}(k,l)$
iFFT
36
$\hat{s}(t)$
Señal de voz con ruido eliminado
Dominio de frecuencia
Dominio temporal

Fig. 1