(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 378 482**

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

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **07290219 .0**
(96) Fecha de presentación: **21.02.2007**
(97) Número de publicación de la solicitud: **1830349**
(97) Fecha de publicación de la solicitud: **05.09.2007**

(54) Título: **Procedimiento de eliminación de ruido de una señal de audio**

(30) Prioridad:
01.03.2006 FR 0601822

(45) Fecha de publicación de la mención BOPI:
13.04.2012

(45) Fecha de la publicación del folleto de la patente:
13.04.2012

(73) Titular/es:
PARROT
174 QUAI DE JEMMAPES
75010 PARIS, FR

(72) Inventor/es:
Pinto, Guillaume

(74) Agente/Representante:
Fàbrega Sabaté, Xavier

DESCRIPCIÓN

Procedimiento de eliminación de ruido de una señal de audio

CONTEXTO DE LA INVENCIÓN

Campo de la invención

La presente invención se refiere a la eliminación de ruido de las señales de audio captadas por un micrófono en un entorno con ruido.

La invención se aplica ventajosamente, pero de modo no limitativo, a las señales de voz captadas por los aparatos telefónicos de tipo “manos-libres” o análogos.

Estos aparatos constan de un micrófono sensible que capta no sólo la voz del usuario, sino igualmente el ruido del entorno, ruido que constituye un elemento perturbador pudiendo llegar, en algunos casos, hasta hacer incomprensibles las palabras del hablante.

Lo mismo sucede si se quieren aplicar técnicas de reconocimiento de voz, en las que es muy difícil operar un reconocimiento de forma sobre palabras sumergidas en un nivel de ruido elevado.

Esta dificultad relacionada con el ruido ambiente es particularmente molesta en el caso de los dispositivos “manos-libres” para vehículos automóviles. En particular, la distancia importante entre el micrófono y el hablante conlleva un nivel relativo de ruido elevado que hace difícil la extracción de la señal útil ahogada por el ruido. Además, el medio con mucho ruido típico del entorno automovilístico 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.

Descripción de la técnica relacionada

Se han propuesto diversas técnicas para reducir el nivel de ruido de la señal captada por un micrófono.

Por ejemplo, el WO-A-98/45997 (Parrot SA) utiliza la presión sobre el pulsador de activación de un teléfono (por ejemplo cuando el conductor quiere responder a una llamada entrante) para detectar el inicio de una señal de voz y considerar que la señal captada antes de presionar era esencialmente una señal de ruido. Esta última señal, memorizada, se analiza para dar un espectro energético medio ponderado del ruido, luego se sustrae de la señal de voz con ruido.

El US-A-5 742 694 describe otra técnica, aplicando un mecanismo de tipo filtro adaptativo predictivo. Este filtro entrega una “señal de referencia” que corresponde a la parte predecible de la señal con ruido y una “señal de error” que corresponde al error de predicción, después atenúa estas dos señales en proporciones variables y las vuelve a combinar para suministrar una señal sin ruido.

El mayor inconveniente de esta técnica de eliminación de ruido reside en la distorsión importante introducida por el filtrado previo, dando en salida una señal muy degradada sobre el plano de la calidad acústica. Además está mal adaptada a las situaciones en las que se necesitaría una eliminación de ruido enérgica con una señal de voz ahogada por un ruido de naturaleza compleja e imprevisible, con características espectrales no estacionarias.

Otras técnicas más, denominadas *beamforming* o *double-phoning*, aplican dos micrófonos distintos. El primero está concebido y colocado para captar principalmente la voz del hablante, mientras que el otro está concebido y colocado para captar una componente de ruido más importante que el micrófono principal. La comparación de las señales captadas permite extraer la voz del ruido ambiente de manera eficaz, y por medios de software relativamente simples.

Esta técnica, basada en un análisis de coherencia espacial de dos señales, presenta no obstante el inconveniente de necesitar dos micrófonos distantes, lo que la relega generalmente con respecto a instalaciones fijas o semifijas y no permite integrarla a un dispositivo preexistente mediante simple añadidura de un módulo software. También presupone que la posición del hablante con respecto a dos micrófonos sea aproximadamente constante, lo que es generalmente el caso en un teléfono de coche utilizado por su conductor. Además, para obtener una eliminación de ruido más o menos satisfactoria, las señales se someten a un filtrado previo importante, lo que presenta, también aquí, el inconveniente de introducir distorsiones que vienen a degradar la calidad de la señal sin ruido restituida.

La invención se refiere a una técnica de eliminación de ruido de las señales de audio captadas por un único micrófono que registra una señal de voz en un entorno con ruido.

Una parte importante de los métodos más eficaces aplicados en los sistemas de un único micrófono se basan en el modelo estadístico establecido por D. Malah e Y. Ephraim en:

[1] Y. Ephraim y D. Malah, Speech Enhancement using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No 6, pp. 1109-1121, Dec. 1984, y
[2] Y. Ephraim y D. Malah, Speech Enhancement using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-33, No 2, pp 443-445, April 1985.

Haciendo la aproximación de que la voz y el ruido son procesos gaussianos no correlacionados y presuponiendo que la potencia espectral del ruido sea un dato conocido, estos dos artículos dan una solución óptima al problema de reducción de ruido descrito más arriba. Esta solución propone cortar la señal con ruido en componentes frecuenciales independientes mediante la utilización de la transformada de Fourier discreta, aplicar una ganancia óptima sobre cada una de estas componentes y después volver a combinar la señal así procesada. Los dos artículos divergen en la elección del criterio de optimalidad. En [1], la ganancia aplicada se denomina *ganancia STSA* y permite minimizar la distancia cuadrática media entre la señal estimada (en la salida del algoritmo) y la señal de voz original (sin ruido). En [2], la aplicación de una ganancia denominada *ganancia LSA* permite en cuanto a ella 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 escogida está en mucha mejor adecuación con el comportamiento del oído humano, y por lo tanto da cualitativamente mejores resultados. En todos los casos, la idea esencial es disminuir la energía de las componentes frecuenciales con mucho ruido aplicándoles una ganancia débil dejando a la vez intactas (mediante la aplicación de una ganancia igual a 1) las que lo son poco o nada.

Aunque es muy atractivo ya que está sostenido por una demostración matemática rigurosa, este procedimiento no puede sin embargo aplicarse solo. En efecto, como se ha indicado más arriba, la potencia espectral del ruido es desconocida e imprevisible *ex ante*. Además, este mismo procedimiento no propone evaluar en qué momentos la voz del hablante está presente en la señal captada. Simplemente se contenta con suponer, o bien que la voz está siempre presente, o bien que está presente una porción fija de tiempo, lo que puede limitar seriamente la calidad de la reducción de ruido.

Por consiguiente, es necesario utilizar otro algoritmo que tenga como función evaluar la potencia espectral del ruido así como los instantes en los que la voz del hablante está presente en la señal bruta captada. Resulta incluso que esta estimación constituye el factor determinante de la calidad de la reducción de ruido operada, siendo el algoritmo de Ephraim y Malah sólo la manera óptima de utilizar la información así obtenida.

Es una solución original a este doble problema de evaluación del ruido y de los instantes de presencia de la señal de voz lo que aporta la presente invención.

Estas dos cuestiones están en realidad intrínsecamente relacionadas. En efecto, supongamos que la señal bruta captada se recorta en tramos de longitudes iguales, de las que se calcula para cada una la transformada de Fourier a corto plazo.

Para una componente frecuencial dada, el conocimiento de los índices de los tramos en los que la voz está ausente permite evaluar la potencia del ruido así como su evolución a lo largo del tiempo en este segmento del espectro. En efecto, basta con medir la energía de la señal bruta cuando la voz está ausente y hacer una media puesta al día continuamente de estas mediciones. Por lo tanto, la cuestión principal es saber cuándo exactamente la voz del hablante está ausente de la señal captada por el micrófono.

Si el ruido es estacionario o pseudoestacionario, este problema se puede resolver fácilmente declarando que la voz está ausente en un segmento de espectro de un tramo dado cuando la energía espectral de los datos para este segmento de espectro no ha evolucionado o ha evolucionado poco con relación a los últimos tramos. Inversamente, se declara que la voz está presente en caso de comportamiento no estacionario.

No obstante, en un entorno real, *a fortiori* un entorno automovilístico en el que más arriba se ha indicado que el ruido conllevaba numerosas características espectrales no estacionarias, este procedimiento es fácilmente cuestionable, en la medida en la que tanto la voz como el ruido pueden presentar comportamientos transitorios. Ahora bien, si se decide conservar todas las componentes transitorias, quedará ruido musical residual en los datos sin ruido; inversamente, si se decide suprimir las componentes transitorias inferiores a un umbral energético dado, entonces las componentes débiles de la voz se borrarán, y estas componentes pueden ser importantes tanto por su contenido informativo como por la inteligibilidad general (distorsión débil) de la señal sin ruido restituida tras procesamiento.

A este respecto, se han propuesto diversos métodos. Entre los más eficaces, se puede citar el descrito por:

[3] I. Cohen y B. Berdugo, Speech Enhancement for Non-Stationary Noise Environments, Signal Processing, Elsevier, Vol. 81, pp. 2403-2418, 2001.

Como frecuentemente en el sector, el procedimiento descrito en este artículo no tiene por objetivo identificar precisamente sobre qué componentes frecuenciales de qué tramos la voz está ausente, sino más bien dar un índice de confianza entre 0 y 1, un valor 1 indicando que la voz está ausente con total seguridad (según el algoritmo) mientras que un valor 0 declara lo contrario. Por su naturaleza, este índice se asimila a la probabilidad de ausencia de la voz *a priori*, es decir, la probabilidad de que la voz esté ausente en una componente frecuencial dada del tramo considerado. Desde luego se trata de una asimilación no rigurosa en el sentido que aunque la presencia de voz es probabilista *ex ante*, la señal captada por el micrófono a cada instante sólo puede pasar por dos estados distintos. Puede, o bien (en el momento considerado) conllevar voz, o bien no contenerla. No obstante, esta asimilación da buenos resultados en la práctica, lo que justifica su utilización. A fin de estimar esta probabilidad de ausencia, Cohen y Berdugo utilizan medias sobre informes señal a ruido *a priori,* utilizados y calculados ellos mismos en el algoritmo de Ephraim y Malah. Estos autores describen igualmente la técnica denominada de ganancia OM-LSA (*Optimally-Modified Log-Spectral Amplitude*), teniendo como objeto mejorar la ganancia LSA por la integración de esta probabilidad de ausencia de la voz.

Esta estimación de la probabilidad *a priori* de ausencia de la voz se revela eficaz, pero depende directamente del modelo estadístico elaborado por Ephraim y Malah y no de un conocimiento *a priori* de los datos.

Para obtener una estimación de la probabilidad de ausencia que sea independiente de este modelo estadístico, Cohen y Berdugo propusieron en:

[4] 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, pp. 233-236, April 2003,

calcular la probabilidad de ausencia a partir de señales captadas por dos micrófonos situados diferentemente, dando señales respectivas en dos vías diferentes, cuya combinación permite obtener una vía denominada de salida y una vía denominada de ruido de referencia. El análisis está basado en la constatación de que las componentes de voz son relativamente más débiles en la vía de ruido de referencia, y que las componentes de ruido transitorio presentan aproximadamente la misma energía en las dos vías. Se determina una probabilidad de presencia de voz para cada segmento de espectro de cada tramo calculando un ratio de energía entre las componentes no estacionarias de las señales respectivas de las dos vías.

Pero, como para las técnicas de *beamforming* o *double-phoning* evocadas más arriba, este procedimiento es bastante incómodo en la medida en que necesita dos micrófonos.

RESUMEN DE LA INVENCIÓN

Uno de los objetivos de la invención es remediar los inconvenientes de los métodos propuestos hasta ahora, gracias a un procedimiento perfeccionado de eliminación de ruido aplicable a una señal de voz considerada aisladamente, en particular una señal captada por un solo micrófono, procedimiento que esté basado en el análisis de la coherencia temporal de las señales captadas.

El punto de partida de la invención reside en la constatación de que la voz presenta generalmente una coherencia temporal superior al ruido y que, por este hecho, es claramente más predecible. Esencialmente, la invención propone utilizar esta propiedad para calcular una señal de referencia en la que la voz se habrá atenuado más que el ruido, aplicando especialmente un algoritmo predictivo que podrá por ejemplo ser del tipo LMS (*Least Mean Squares*, método de mínimos cuadrados). Esta señal de referencia derivada de la señal de voz de la que hay que eliminar el ruido se podrá utilizar de manera comparable a la de la señal del segundo micrófono de las técnicas de *beam-forming* de dos vías, por ejemplo de las técnicas similares a las de Cohen y Berdugo [4, citado anteriormente]. El cálculo de un ratio entre los niveles de energía respectivos de la señal original y de la señal de referencia así obtenido permitirá discriminar entre las componentes de voz y los ruidos parásitos no estacionarios, y suministrará una estimación de la probabilidad de presencia de voz de manera independiente de todo modelo estadístico.

En otras palabras, la técnica propuesta por la invención aplica una “sustracción inteligente” que implica, tras una predicción lineal operada en base a las muestras tratadas de la señal original (y no de una señal previamente filtrada, por consiguiente degradada), un reajuste de fase entre la señal original y la señal predicha.

El rendimiento de la técnica de la invención se revela, en la práctica, suficiente como para asegurar una eliminación de ruido extremadamente eficaz directamente sobre la señal original, liberándose de distorsiones introducidas por una cadena de filtrado previo, convertida en inútil.

Más precisamente, la presente invención propone, para la eliminación de ruido de una señal de audio original con ruido que conlleva una componente de voz combinada a una componente de ruido que conlleva ella misma una

componente de ruido transitoria y una componente de ruido pseudoestacionaria, operar un análisis de coherencia temporal de la señal con ruido por las etapas de:

a) determinación de una señal de referencia por aplicación a la señal con ruido de un procesamiento propio para atenuar de forma más importante las componentes de voz que las componentes de ruido de esta señal con ruido, comprendiendo dicho procesamiento: (a1) la aplicación de un algoritmo de predicción lineal adaptativo que opera sobre una combinación lineal de las muestras anteriores de la señal con ruido, y (a2) la determinación de dicha señal de referencia por una sustracción, con compensación del desfase, entre la señal original con ruido, no filtrada y la señal entregada por el algoritmo de predicción lineal;
b) determinación de una probabilidad de presencia/ausencia de voz *a priori* a partir de los niveles de energía respectivos en el dominio espectral de la señal con ruido y de la señal de referencia; y
c) utilización de esta probabilidad de ausencia de voz *a priori* para estimar un espectro de ruido y derivar de la señal con ruido una estimación sin ruido de la señal de voz.

La señal de referencia se puede determinar en especial por aplicación en la etapa a2) de una relación del tipo:

$$Ref(k,l) = X(k,l) - X(k,l)\frac{|Y(k,l)|}{|X(k,l)|}$$

donde *X(k,l)* e *Y(k,l)* son las transformadas de Fourier a corto plazo de cada segmento de espectro *k* de cada tramo *l*, respectivamente de la señal original con ruido y de la señal entregada por el algoritmo de predicción lineal.

El algoritmo predictivo es ventajosamente un algoritmo adaptativo recursivo del tipo método de mínimos cuadrados LMS.

La etapa b) comprende ventajosamente la aplicación de un algoritmo de estimación de la energía de la componente de ruido pseudoestacionaria en la señal de referencia y en la señal con ruido, en especial un algoritmo de tipo de cálculo recursivo del promedio controlado por mínimos MRCA como se describe en:

> [5] I. Cohen y B. Berdugo, Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement, IEEE Signal Processing Letters, Vol. 9, No 1, pp 12-15, Jan. 2002,

La etapa c) comprende ventajosamente la aplicación de un algoritmo de ganancia variable función de la probabilidad de presencia/ausencia de voz, en espacial un algoritmo de tipo ganancia de amplitud log-espectral modificado optimizado OM-LSA.

DESCRIPCIÓN SUMARIA DE LOS DIBUJOS

A continuación se va a describir un ejemplo de aplicación de la invención, con referencia a los dibujos adjuntos en los que las mismas referencias numéricas designan de una figura a otra, elementos idénticos o funcionalmente semejantes.

> La figura 1 es un diagrama esquemático que ilustra las diferentes operaciones efectuadas por un algoritmo de eliminación de ruido conforme al procedimiento de la invención
> La figura 2 es un diagrama esquemático que ilustra más particularmente el algoritmo predictivo LMS adaptativo.

DESCRIPCIÓN DETALLADA DE LA FORMA DE REALIZACIÓN PREFERENTE

La señal de la que se desea eliminar el ruido es una señal numérica muestreada *x(n)*, en la que *n* designa el número de la muestra (*n* es por lo tanto la variable temporal).

La señal captada *x(n)* es una combinación de una señal de voz *s(n)* y de un ruido sobreañadido, no correlacionado, *d(n):*

$$x(n) = s(n) + d(n)$$

Este ruido *d(n)* tiene dos componentes independientes, a saber una componente transitoria $d_t(n)$ y una componente pseudoestacionaria $d_{ps}(n)$:

$$d(n) = d_t(n) + d_{ps}(n)$$

Como se ilustra en la figura 1, la señal con ruido *x(n)* se aplica en la entrada de un algoritmo LMS predictivo esquematizado por el bloque 10, incluyendo la aplicación de retardos apropiados 12. El funcionamiento de este algoritmo LMS se describirá más abajo, con referencia a la figura 2.

A continuación se calcula la transformada de Fourier a corto plazo de la señal captada *x(n)* (bloque 16), así como de la señal *y(n)* entregada por el algoritmo LMS predictivo (bloque 14). A partir de estas dos transformadas se calcula una señal de referencia (bloque 18), que constituye una de las variables de entrada de un algoritmo de cálculo de la

probabilidad de ausencia de voz (bloque 24). Paralelamente, la transformada de la señal con ruido *x(n)*, resultante del bloque 16, se aplica igualmente al algoritmo de cálculo de probabilidad.

Los bloques 20 y 22 estiman el ruido pseudoestacionario de la señal de referencia y de la transformada de la señal con ruido es estimada, y el resultado es igualmente aplicado al algoritmo de cálculo de probabilidad.

El resultado del cálculo de probabilidad de ausencia de voz, así como la transformada de la señal con ruido, se aplican en la entrada de un algoritmo de procesamiento de ganancia OM-LSA (bloque 26), cuyo resultado se somete a una transformación inversa de Fourier (bloque 28) para dar una estimación de la voz sin ruido.

A continuación se van a describir con más detalle las diferentes fases de este procesamiento.

El algoritmo predictivo LMS (bloque 10) se esquematiza en la figura 2.

En la medida en que las señales en presencia son globalmente no estacionarias pero localmente pseudoestacionarias, se puede utilizar ventajosamente un sistema adaptativo, que podrá tener en cuenta variaciones de energía de la señal en el tiempo y converger hacia los diversos locales óptimos.

Esencialmente, si se aplican retardos sucesivos Δ, la predicción lineal *y(n)* de la señal *x(n)* es una combinación lineal de las muestras anteriores $\{x(n - \Delta - i + 1)\}_{1 \le i \le M}$:

$$y(n) = \sum_{i=1}^{M} w_i x(n - \Delta - i + 1)$$

que minimiza el error cuadrático medio del error de predicción:

$$\epsilon(n) = x(n) - y(n)$$

La minimización consiste en encontrar:

$$\min_{w_1, w_2, \ldots, w_M} E\left[x(n) - \sum_{i=1}^{M} w_i x(n - \Delta - i + 1)\right]^2$$

Para resolver este problema, es posible utilizar un algoritmo LMS, que es un algoritmo en sí mismo conocido, descrito por ejemplo en:

[6] B. Widrow, Adaptative Filter, Aspect of Network and System Theory, R. E. Kalman and N. De Claris (Eds). New York: Holt, Rinehart and Winston, pp. 563-587, 1970, y
[7] B. Widrow y al., Adaptative Noise Cancelling: Principles and Applications, Proc. IEEE, Vol. 63, No 12 pp. 1692-1716, Dec 1975.

Se puede definir un procedimiento recursivo de adaptación de las ponderaciones.

$$w_i(n + 1) = w_i(n) + 2\mu\epsilon(n)x(n - \Delta - i + 1)$$

siendo μ una constante de ganancia que permite ajustar la velocidad y la estabilidad de la adaptación.

Se podrán encontrar indicaciones generales sobre estos aspectos del algoritmo LMS en:

[8] B. Widrow y S. Stearns, Adaptative Signal Processing, Prentice-Hall Signal Processing Series, Alan V. Oppenheim Series Editor, 1985.

Se puede demostrar que tal predicción lineal adaptativa permite discriminar eficazmente entre ruido y voz ya que las muestras que contienen la voz se predecirán mucho mejor (errores cuadráticos más pequeños entre la predicción y la señal bruta) que los que sólo contienen ruido.

Más precisamente, las señales respectivas *x(n)* e *y(n)* (señal de voz con ruido y predicción lineal) se recortan en tramos de longitudes idénticas, y su transformada de Fourier a corto plazo (marcadas respectivamente *X* e *Y*) se calcula para cada tramo. Para evitar los efectos de los errores de precisión, el algoritmo prevé un recubrimiento del 50% entre tramos consecutivos, y las muestras se multiplican por los coeficientes de la ventana de Hanning de manera que la suma de los tramos pares e impares corresponde a la señal de origen propiamente dicha. Para el segmento de espectro *k* de un tramo *l* par, se tiene:

$$X(k,l) = \sum_{p=1}^{R} h(p)x(Rl+p)e^{-j2\pi\frac{pk}{R}}$$

Y para el segmento de espectro *k* de un tramo *l* impar:

$$X(k,l) = \sum_{p=1}^{R} h(p)x(\frac{R}{2}l+p)e^{-j2\pi\frac{pk}{R}}$$

siendo *h* la ventana de Hanning.

Una primera posibilidad consiste en definir la señal de referencia tomando la transformada de Fourier del error de predicción:

$$\hat{\epsilon}(k,l) = X(k,l) - Y(k,l)$$

No obstante, se constata en la práctica un cierto desfase entre *X* e *Y* debido a una convergencia imperfecta del algoritmo LMS, impidiendo una buena discriminación entre voz y ruido. Por consiguiente, se prefiere adoptar para la señal de referencia otra definición que compense este desfase, a saber:

$$Ref(k,l) = X(k,l) - X(k,l)\frac{|Y(k,l)|}{|X(k,l)|}$$

Se supone que la energía espectral de la señal de referencia se puede describir bajo la forma:

$$E[Ref(k,l)]^2 = E[S(k,l)]^2\alpha_S(k) + E[D_t(k,l)]^2\alpha_{D_t}(k) + E[D_{ps}(k,l)]^2\alpha_{D_{ps}}(k)$$

donde

$$\alpha_S(k) < \alpha_{D_t}(k) < \alpha_{D_{ps}}(k)$$

representan la atenuación en la señal de referencia de las tres señales en cada segmento de espectro.

La etapa siguiente consiste en entregar una estimación *q(k,l)* de la probabilidad de ausencia de voz en la señal con ruido:

$$q(k,l) = Pr\{H_0(k,l)\}$$

$H_0(k,l)$ indicando la ausencia de voz (y $H_1(k,l)$ la presencia de voz) en el $k^{ésimo}$ segmento de espectro del $l^{ésimo}$ tramo.

La discriminación entre ruido transitorio y voz se puede operar mediante una técnica comparable a la de Cohen y Berdugo (5, citada anteriormente). Más precisamente, el algoritmo de la invención evalúa un ratio de las energías transitorias en las dos vías, dado por:

$$\Omega(k,l) = \frac{SX(k,l) - MX(k,l)}{SRef(k,l) - MRef(k,l)}$$

siendo *S* una estimación suavizada de la energía instantánea:

$$SX(k,l) = SX(k,l-1) + \sum_{i=-\omega}^{\omega} b(i)|X(k,l)|^2$$

siendo *b* una ventana en el dominio temporal y siendo *M* un estimador de la energía pseudoestacionaria, que se puede obtener por ejemplo por un método MCRA (*Mínima Controlled Recursive Averaging*) del mismo tipo que el descrito por Cohen y Berdugo [5, citado anteriormente] (no obstante existen varias alternativas en la literatura).

En presencia de voz pero en ausencia de ruido transitorio, este ratio vale aproximadamente:

$$\Omega(k,l) = \frac{1}{\alpha_{D_t}(k)} = \Omega_{max}(k)$$

Inversamente, en ausencia de voz pero en presencia de ruidos transitorios:

$$\Omega(k,l) = \frac{1}{\alpha g(k)} = \Omega_{min}(k)$$

Si se supone que en general:

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

un procedimiento de estimación de *q(k,l)* se da por el algoritmo en metalenguaje siguiente:

Para cada tramo l y para cada segmento de espectro k,

(i) Calcular SX(k,l), MX(k,l), SRef(k,l) y MRef(k,l). Ir a (ii)
(ii) Si SX(k,l) > L_xMX(k,l) (detección de transitorios en la vía de voz con ruido), entonces ir a (iii) si no

$$q(k,l) = 1$$

(iii) Si SRef(k,l) > L_{Ref}MRef(k,l) (detección de transitorios en la vía de referencia), entonces ir a (iv) si no

$$q(k,l) = 0$$

(iv) Calcular Ω(k,l), ir a (v)
(v) Calcular:

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

Las constantes L_x y L_{Ref} son umbrales de detección de transitorios. $\Omega_{min}(k)$ y $\Omega_{max}(k)$ son los límites superior e inferior para cada segmento de espectro. Estos diversos parámetros se escogen de manera que correspondan con situaciones típicas, próximas a la realidad.

La etapa siguiente (correspondiente al bloque 26 de la figura 1) consiste en operar la eliminación de ruido propiamente dicha (refuerzo de la componente de voz). El estimador que se acaba de describir se aplicará al modelo estadístico descrito por Ephraim y Malah [2, citado anteriormente], que supone que el ruido y la voz en cada segmento de espectro son procesos gaussianos independientes de varianzas respectivas $\lambda_x(k,l)$ y $\lambda_d(k,l)$.

Esta etapa puede aplicar ventajosamente el algoritmo de ganancia OM-LSA *(Optimally Modified Log-Spectral Amplitude Gain)* descrito por Cohen y Berdugo [3, citado anteriormente]. La relación señal/ruido *a priori* se define por:

$$\xi(k,l) = \frac{\lambda_x(k,l)}{\lambda_d(k,l)}$$

La relación señal/ruido *a posteriori* se define por:

$$\gamma(k,l) = \frac{|X(k,l)|^2}{\lambda_d(k,l)}$$

La probabilidad condicional de presencia de la señal es:

$$p(k,l) = Pr(H_1(k,l)|X(k,l))$$

Con la hipótesis gaussiana y los parámetros anteriores, viene:

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))exp(-v(k,l))\right\}^{-1}$$

con:

$$v(k,l) = \frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}$$

La óptima estimación de la voz con eliminación de ruido *S(k,l)* se da por:

$$\hat{S}(k,l) = G_{H_1}(k,l)^{p(k,l)} G_{min}^{1-p(k,l)} X(k,l)$$

siendo G_{H1} la ganancia en la hipótesis en la que la voz está presente, que se define por:

$$G_{H_1}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} exp\left(\frac{1}{2}\int_{v(k,l)}^{\infty} \frac{e^{-t}}{t} dt\right)$$

La ganancia G_{min} en la hipótesis de ausencia de voz es un límite inferior para la reducción del ruido, a fin de limitar la distorsión de la voz.

La fórmula clásica de estimación de la relación señal/ruido *a priori* es:

$$\hat{\xi}(k,l) = aG_{H_1}^2(k,l-1)\gamma(k,l-1) + (1-a)\max(\gamma(k,l)-1,0)$$

La estimación de la energía del ruido se da por:

$$\hat{\lambda}_d(k,l+1) = \tilde{\alpha}_d(k,l)\hat{\lambda}_d(k,l) + \beta(1-\tilde{\alpha}_d(k,l))|X(k,l)|^2$$

El parámetro de suavizado $\tilde{a}_d$ evoluciona entre un límite inferior a_d y 1, en función de la probabilidad de presencia condicional:

$$\tilde{\alpha}_d(k,l) = \alpha_d + (1-\alpha_d)p(k,l)$$

siendo β un factor de sobreestimación que compensa el sesgo en ausencia de señal.

La señal obtenida después de este procesamiento se somete a una transformada de Fourier inversa (bloque 28) para dar la estimación final de la voz con eliminación de ruido.

El algoritmo de la presente invención resulta particularmente eficaz en los entornos ruidosos, a la vez parasitados por ruidos mecánicos, vibraciones, etc., así como por ruidos musicales, situaciones características encontradas en el habitáculo de un coche. Los espectrogramas muestran que la atenuación del ruido no es sólo eficaz, sino que se realiza sin distorsión notable de la voz tras la eliminación de ruido.

REIVINDICACIONES

1. Un procedimiento de procesamiento de una señal de audio, para la eliminación de ruido de una señal original con ruido que consta de una componente de voz combinada con una componente de ruido, esta componente de ruido comprende ella misma una componente de ruido transitoria y una componente de ruido pseudoestacionaria,
caracterizado por que este procedimiento es un procedimiento de análisis de coherencia temporal de la señal con ruido muestreada que comprende las etapas de:

a) determinación de una señal de referencia por aplicación a la señal con ruido de un procesamiento (10,18) propio para atenuar de manera más importante las componentes de voz que las componentes de ruido de esta señal con ruido, dicho procesamiento comprendiendo:

a1) la aplicación de un algoritmo de predicción lineal adaptativo que opera sobre una combinación lineal de las muestras anteriores de la señal con ruido, y
a2) la determinación de dicha señal de referencia por una sustracción, con compensación del desfase, entre la señal con ruido original, no filtrada previamente y la señal entregada por el algoritmo de predicción lineal;

b) determinación (24) de una probabilidad de presencia/ausencia de voz *a priori* a partir de los niveles de energía respectivos en el dominio espectral de la señal con ruido y de la señal de referencia; y

c) utilización de esta probabilidad de ausencia de voz *a priori* para estimar un espectro de ruido y derivar (26) de la señal con ruido una estimada con eliminación de ruido de la señal de voz.

2. El procedimiento de la reivindicación 1, en el que dicha señal de referencia se determina por aplicación en la etapa a2) de una relación del tipo:

$$Ref(k, l) = X(k, l) - X(k, l)\frac{|Y(k, l)|}{|X(k, l)|}$$

donde *X(k,l)* e *Y(k,l)* son las transformadas de Fourier a corto plazo de cada segmento de espectro *k* de cada tramo *l*, respectivamente de la señal original con ruido y de la señal entregada por el algoritmo de predicción lineal.

3. El procedimiento de la reivindicación 1, en el que el algoritmo de predicción lineal (10) es un algoritmo del tipo método de mínimos cuadrados LMS.

4. El procedimiento de la reivindicación 1, en el que el algoritmo de predicción lineal (10) es un algoritmo adaptativo recursivo.

5. El procedimiento de la reivindicación 1, en el que la etapa b) comprende la aplicación de un algoritmo de estimación de la energía de la componente de ruido pseudoestacionaria en la señal de referencia y en la señal con ruido.

6. El procedimiento de la reivindicación 5, en el que el algoritmo de estimación de la energía de la componente de ruido pseudoestacionaria es un algoritmo de tipo de cálculo recursivo del promedio controlado por mínimos MRCA.

7. El procedimiento de la reivindicación 1, en el que la etapa c) comprende la aplicación de un algoritmo de ganancia variable función de la probabilidad de presencia/ausencia de voz.

8. El procedimiento de la reivindicación 7, en el que el algoritmo de ganancia variable es un algoritmo de tipo ganancia de amplitud log-espectral modificado optimizado OM-LSA.

FIG_1

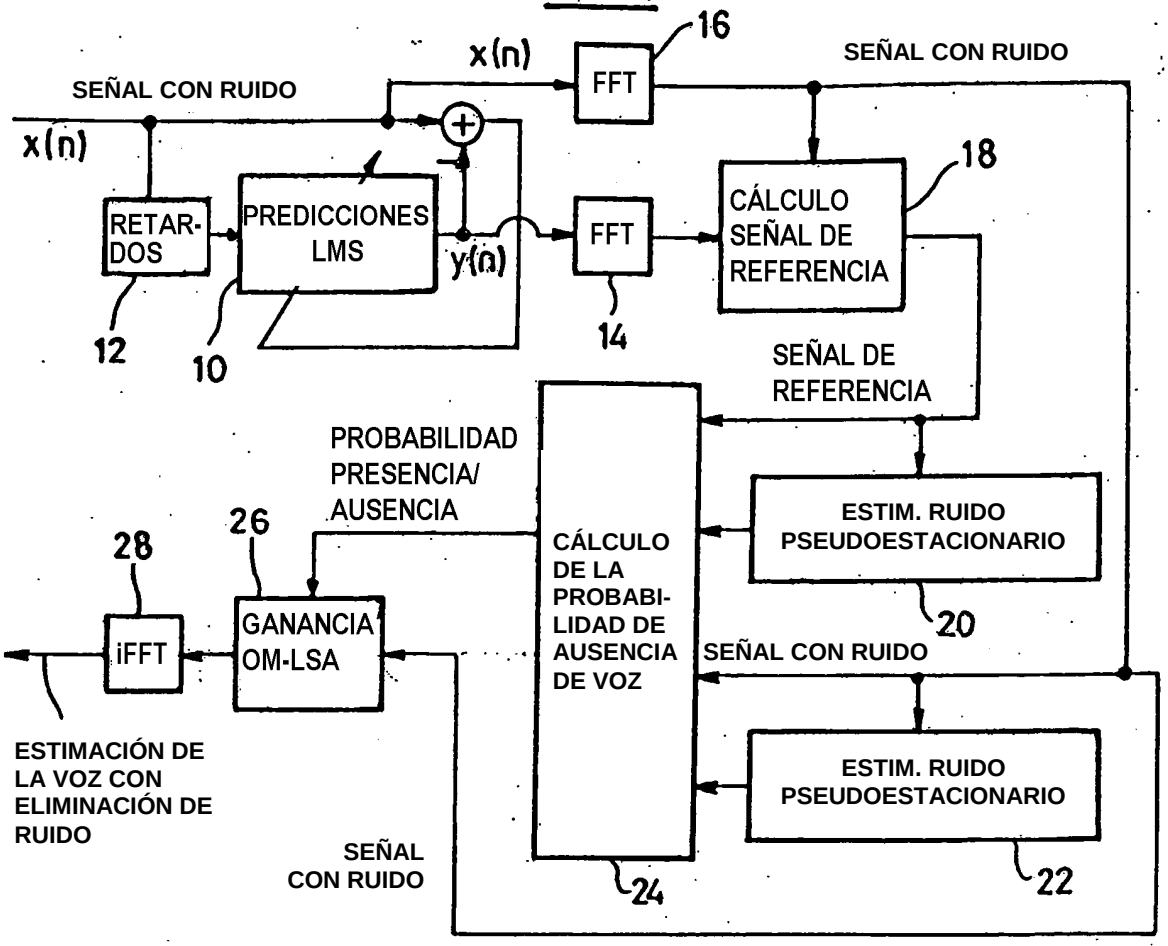
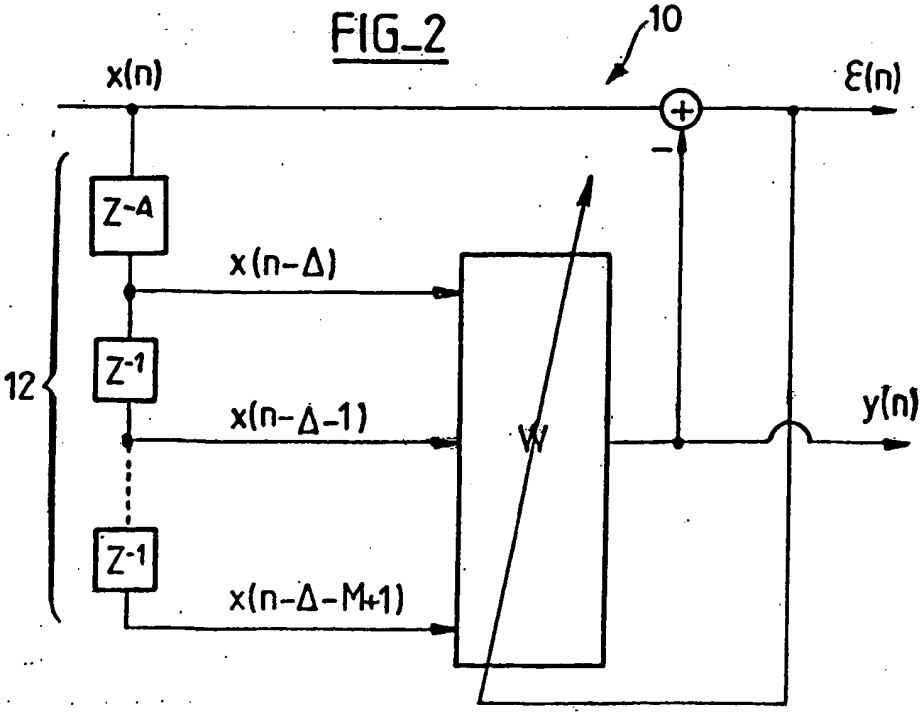

SEÑAL CON RUIDO
x(n)
FFT
16
SEÑAL CON RUIDO
x(n)
RETAR-
DOS
PREDICCIONES
LMS
y(n)
FFT
CÁLCULO
SEÑAL DE
REFERENCIA
18
12
10
14
SEÑAL DE
REFERENCIA
PROBABILIDAD
PRESENCIA/
AUSENCIA
ESTIM. RUIDO
PSEUDOESTACIONARIO
20
28
26
CÁLCULO
DE LA
PROBABI-
LIDAD DE
AUSENCIA
DE VOZ
iFFT
GANANCIA
OM-LSA
SEÑAL CON RUIDO
ESTIMACIÓN DE
LA VOZ CON
ELIMINACIÓN DE
RUIDO
ESTIM. RUIDO
PSEUDOESTACIONARIO
SEÑAL
CON RUIDO
24
22

FIG_2


10
x(n)
ε(n)
Z-Δ
x(n-Δ)
12
Z-1
x(n-Δ-1)
W
y(n)
Z-1
x(n-Δ-M+1)