(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 294 506**

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

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Número de solicitud europea: **04741579 .9**
(86) Fecha de presentación : **14.05.2004**
(87) Número de publicación de la solicitud: **1745468**
(87) Fecha de publicación de la solicitud: **24.01.2007**

(54) Título: **Reducción de ruido para el reconocimiento automático del habla.**

(73) Titular/es: **Loquendo S.p.A.**
Via Valdellatorre, 4
10149 Torino, IT

(45) Fecha de publicación de la mención BOPI:
01.04.2008

(72) Inventor/es: **Gemello, Roberto y**
Mana, Franco

(45) Fecha de la publicación del folleto de la patente:
01.04.2008

(74) Agente: **Ponti Sales, Adelaida**

DESCRIPCIÓN

Reducción de ruido para el reconocimiento automático del habla.

Campo técnico de la invención

La presente invención se refiere de forma general a la reducción de ruido para el reconocimiento automático del habla, y en concreto a un procedimiento y sistema de reducción de ruido basado en la técnica de atenuación espectral, y a un sistema de reconocimiento automático del habla.

Antecedentes técnicos

La figura 1 muestra un diagrama de bloques de fuentes usuales de degradación del habla. Como se puede apreciar, el habla del hablante que se desea (bloque 10) se degrada debido al ruido ambiental, en concreto voces de otros hablantes cercanos (bloque 20) y ruido de fondo (bloque 30), y debido al ruido y la distorsión del canal de comunicación (bloques 40 y 50). Las técnicas de reducción de ruido (bloque 60) para el reconocimiento automático del habla (bloque 70) pueden reducir el ruido de fondo y el ruido de canal (casi estacionarios), mientras que el ruido no estacionario y las voces interferentes son mucho más difíciles de eliminar.

La figura 2 muestra un diagrama de bloques de un sistema automático de reconocimiento del habla. Como se puede apreciar, el habla ruidosa a reconocer se introduce en un bloque 100 de análisis espectral de tiempo corto (FFT enventanada) que genera espectros de tiempo corto que se introducen a su vez en un bloque reductor de ruido 110. Los espectros de tiempo corto sin ruido se introducen en una etapa de entrada 120 de RASTA-PLP, que a su vez indica a la salida la energía total de la señal de habla, los coeficientes de cepstrum, y las derivadas primera y segunda de la energía total y de los coeficientes de cepstrum, introduciéndose todos ellos en un bloque 130 de reconocimiento automático del habla.

La etapa de entrada 120 de RASTA-PLP implementa una técnica que se conoce como "RelaAtive SpecTrAl Technique", que es un mejora sobre el procedimiento de PLP (predicción perceptual lineal) tradicional y que consiste en un filtrado especial de los diferentes canales de frecuencia de un analizador PLP. El filtrado previo se realiza para hacer que el análisis del habla sea menos sensible a los cambios lentos o los factores de estado estacionario del habla. El procedimiento RASTA substituye al espectro de plazo corto de banda crítica convencional del PLP e introduce una estimación espectral menos sensible. Para una descripción más detallada de un procesado RASTA, se puede consultar la referencia de H. Hermansky y N. Morgan, "RASTA Processing of Speech", IEEE Transactions on Speech and Audio Processing, volumen 2, número 4, octubre de 1994.

El bloque de reducción de ruido 110 realiza una estimación del ruido ambiental 112 basada en los espectros de tiempo corto y a continuación una reducción de ruido ambiental 114 basada en los espectros de tiempo corto y el ruido estimado, utilizando una técnica que se denomina de "substracción espectral" o una técnica que se denomina de "atenuación espectral".

Las técnicas arriba mencionadas se describirán con más detalle a continuación, donde se indicará el espectro de potencia del habla ruidosa como $|Y_k(m)|^2$, el espectro de potencia del habla limpia como $|X_k(m)|^2$, el espectro de potencia del ruido aditivo como $|D_k(m)|^2$, y la estimación de una cantidad por medio del símbolo "^", y donde k es el índice de las líneas espectrales de los espectros y m es el índice de las ventanas de tiempo dentro de las cuales se procesa el habla ruidosa para la reducción del ruido.

La técnica de substracción espectral se describe en N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", IEEE Transactions on Speech and Audio Processing, volumen 7, número 2, marzo de 1999, donde se trata el problema de la reducción del ruido para el reconocimiento del habla y se describe la utilización de un factor de sobreestimación o sobresubstracción y un factor de fondo espectral.

En concreto, la técnica de substracción espectral se basa en el principio de reducir el ruido por medio de substraer una estimación del espectro de potencia del ruido aditivo $|\hat{D}_k(m)|^2$ del espectro de potencia del habla ruidosa $|Y_k(m)|^2$, obteniéndose de esta forma una estimación $|\hat{X}_k(m)|^2$ del espectro de potencia del habla limpia:

$$|\hat{X}_k(m)|^2 = \begin{cases} |Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2 & \text{si } |Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2 > \beta(m)|Y_k(m)|^2 \\ \beta(m)|Y_k(m)|^2 & \text{en otro caso} \end{cases} \quad (1)$$

donde $\alpha(m)$ es el factor de sobreestimación de ruido, $\beta(m)$ es el factor de fondo de espectro.

En concreto, el espectro de ruido residual consiste en crestas y valles con aparición aleatoria, y el factor de sobreestimación $\alpha(m)$ y el factor de fondo de espectro $\beta(m)$ se han introducido para reducir las excursiones espectrales.

En detalle, el factor de sobreestimación $\alpha(m)$ se ha introducido para "sobreestimar" el espectro de ruido, es decir, en otras palabras el factor de sobreestimación $\alpha(m)$ substrae una sobreestimación del ruido a lo largo del espectro completo, mientras que el factor de fondo espectral $\beta(m)$ evita que las líneas espectrales de la estimación del espectro de potencia $|\hat{X}_k(m)|^2$ del habla limpia caigan por debajo de un límite inferior ($\beta(m)|Y_k(m)|^2$), "rellenando" de esta forma los valles profundos que rodean a los picos estrechos (del espectro mejorado). De hecho, ocasionalmente se pueden producir estimaciones negativas del espectro de potencia mejorado y en tales casos, las líneas espectrales negativas se llevan a cero o a algún valor mínimo (de fondo). Reducir las excursiones espectrales de los picos de ruido, en comparación con cuando se establecen en cero los componentes negativos, reduce la cantidad de ruido musical. Esencialmente por medio de reinsertar el ruido de banda ancha (fondo de ruido), los remanentes de los picos de ruido se "enmascaran" por parte de los componentes vecinos de magnitud comparable.

Una variante de esta técnica se conoce como "técnica de substracción espectral de Wiener", que es similar a la anterior pero que se deriva de la teoría de filtrado óptimo. La estimación $|\hat{X}_k(m)|^2$ del espectro de potencia del habla limpia es la siguiente:

$$|\hat{X}_k(m)|^2 = \begin{cases} \dfrac{\left[|Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2\right]^2}{|Y_k(m)|^2} & \text{si } |Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2 > \beta(m)|Y_k(m)|^2 \\ \beta(m)|Y_k(m)|^2 & \text{en otro caso} \end{cases} \quad (2)$$

Una mejora sobre las técnicas de substracción espectral se describe en V. Schless, F. Class, "SNR-Dependent Flooring and Noise Overestimation for Joint Application of Spectral Substraction and Model Combination", ICSLP 1998, donde se propone hacer que el factor de sobreestimación de ruido $\alpha(m)$ y el factor de fondo espectral $\beta(m)$ sean funciones de la relación de señal a ruido global SNR(m).

La técnica de atenuación espectral, por el contrario, se basa en el principio de suprimir el ruido por medio de aplicar una regla de supresión, o una ganancia de valor real no negativa G_k, a cada línea espectral k del espectro de magnitud $|Y_k(m)|$ del habla ruidosa, para calcular una estimación $|\hat{X}_k(m)|$ del espectro de magnitud del habla limpia según la siguiente fórmula:

$$|\hat{X}_k(m)| = G_k(m)|Y_k(m)|$$

Se han propuesto muchas reglas de supresión, y probablemente una de las reglas más importantes es la que se denomina regla logarítmica de atenuación espectral de Ephraim-Malah, que se describe en Y. Ephraim y D. Malah, "Speech Enhancement Using a Minimum Min-Square Error Log-Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing, volumen ASSP-33, número 2, páginas 443-445, 1985.

La ganancia de Ephraim-Malah $G_k(m)$ se define como:

$$G_k(m) = \frac{\xi_k(m)}{1 + \xi_k(m)} \exp\left(\frac{1}{2}\int_{\nu_k}^{\infty} \frac{e^{-t}}{t}\, dt\right) \quad (3)$$

donde:

- $\xi_k(m)$ es una relación de señal a ruido *a priori* relativa a la línea espectral k-ésima y se define de la siguiente forma:

$$\xi_k(m) = \frac{|X_k(m)|^2}{|D_k(m)|^2} \quad (4)$$

- $\nu_k(m)$ se define como:

$$\nu_k(m) = \frac{\xi_k(m)}{1 + \xi_k(m)} \gamma_k(m) \quad (5)$$

- $\gamma_k(m)$ es una relación de señal a ruido que se denomina *a posteriori* relativa a la línea espectral k-ésima y se define de la siguiente forma:

$$\gamma_k(m) = \frac{|Y_k(m)|^2}{|D_k(m)|^2} \qquad (6)$$

El cálculo de la relación de señal a ruido *a posteriori* $\gamma_k(m)$ requiere el conocimiento del espectro de potencia del ruido aditivo $|D_k(m)|^2$, que no se encuentra disponible. Se puede obtener una estimación $|\hat{D}_k(m)|^2$ del espectro de potencia del ruido aditivo con un estimador de ruido como se describe en H. G. Hirch, C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition", ICASSP 1995, páginas 153-156.

Por tanto, se puede calcular una estimación $\hat{\gamma}_k(m)$ de la relación de señal a ruido *a posteriori* de la siguiente forma:

$$\hat{\gamma}_k(m) = \frac{|Y_k(m)|^2}{|\hat{D}_k(m)|^2} \qquad (7)$$

El cálculo de la relación de señal a ruido *a priori* $\xi_k(m)$ requiere el conocimiento del espectro de potencia $|X_k(m)|^2$ del habla limpia, el cual no se encuentra disponible. Se puede calcular una estimación $\hat{\xi}_k(m)$ de la relación de señal a ruido *a priori* por medio de la utilización de una aproximación dirigida a la decisión como se describe en 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, volumen ASSP-32, número 6, páginas 1109-1121, 1984, y de la forma siguiente:

$$\hat{\xi}_k(m) = \eta(m)\frac{|\hat{X}_k(m-1)|^2}{|\hat{D}_k(m-1)|^2} + [1-\eta(m)]\max[0, \hat{\gamma}_k(m)-1], \eta(m) \in [0,1) \qquad (8)$$

donde $\eta(m)$ es un coeficiente de peso para ponderar adecuadamente los dos términos de la fórmula.

La ganancia de Ephraim-Malah $G_k(m)$ se puede calcular entonces como función de la estimación $\hat{\xi}_k(m)$ de la relación de señal a ruido *a priori* y de la estimación $\hat{\gamma}_k(m)$ de la relación de señal a ruido *a posteriori* según la fórmula (3).

En US-A-2002/0002455 se describe una aplicación de la técnica de atenuación espectral, que se refiere a un sistema de mejora del habla que recibe un habla ruidosa caracterizada por una amplitud espectral que se extiende en una pluralidad de cajones y que produce un habla mejorada por medio de modificar la amplitud espectral del habla ruidosa sin afectar a la fase de la misma. En concreto, el sistema de mejora del habla comprende un estimador de núcleo que aplica al habla ruidosa uno de un primer conjunto de ganancias para cada cajón de frecuencia; un módulo de adaptación de ruido que segmenta el habla ruidosa en cuadros que contienen solamente ruido y solamente señal, mantiene una estimación actual del espectro de ruido y una estimación de la probabilidad de ausencia de señal en cada cajón de frecuencia; y un estimador de relación de señal a ruido que mide la relación de señal a ruido *a posteriori* y estima una relación de señal a ruido *a priori* basándose en la estimación de ruido. Cada uno del primer conjunto de ganancias se basa en una relación de señal a ruido *a priori*, así como en la probabilidad de ausencia de señal en cada cajón y en un nivel de agresividad de la mejora del habla. Un módulo de decisión suave calcula un segundo conjunto ganancias que se basa en una relación de señal a ruido *a posteriori* y en una relación de señal a ruido *a priori*, y la probabilidad de ausencia de señal en cada cajón de frecuencia.

En WO-A-01/52242 se describe otra aplicación de las técnicas de atenuación espectral, que se refiere a un esquema de substracción espectral de banda múltiple que se puede aplicar a una variedad de sistemas de comunicación por habla, como soportes a la audición, sistemas de acceso público, sistemas de teleconferencia, sistemas de control por voz, o sistemas altavoces, y que comprende una arquitectura una arquitectura de filtro de banda múltiple, detección de potencia de ruido y de señal, y función de ganancia para la reducción del ruido. La función de ganancia para la reducción de ruido consiste en una función de escala de ganancia y una función de atenuación máxima que proporcionan una cantidad de ganancia predeterminada como función de la relación de señal a ruido y del ruido. La función de escala de ganancia es una función a tramos lineal de tres segmentos, y los tres tramos lineales de la función de escala de ganancia comprenden un primer tramo que proporciona una expansión máxima hasta un primer punto de codo para una reducción de ruido máxima, un segundo tramo que proporciona una expansión menor hasta un segundo punto de codo para una reducción de ruido menor, y un tercer tramo que proporciona una expansión mínima o nula para señales de entrada con una relación de señal a ruido alta para minimizar la distorsión. La función de atenuación máxima puede ser una constante o ser igual a la envolvente del ruido estimada. Cuando se utiliza en aplicaciones de soporte a la audición, la función de ganancia de reducción de ruido se combina con la función de ganancia de compensación de pérdida de la audición inherente al proceso de asistencia a la audición.

El reconocimiento automático del habla que utiliza los procedimientos de reducción de ruido conocidos arriba descritos se encuentra afectado por algunos problemas técnicos que evitan que sea realmente efectivo. En concreto, la técnica de substracción espectral y la técnica de substracción espectral de Wiener se ven afectadas por el denominado "ruido musical", que se introduce en el espectro de potencia $|X_k(m)|^2$ del habla limpia por el nivel de fondo arriba mencionado, según el cual los valores negativos se establecen en un valor de fondo $\beta(m)|Y_k(m)|$ para evitar la ocurrencia de resultados negativos de substracción. En concreto, el nivel de fondo introduce discontinuidades en el espectro que se perciben como ruidos musicales molestos y que degradan el rendimiento de un sistema de reconocimiento automático del habla.

La técnica de atenuación espectral que implementa la regla de atenuación de Ephraim-Malah es una técnica muy buena para la denominada mejora del habla, es decir la reducción del ruido para un oyente humano, pero introduce cierta distorsión espectral en partes de la voz que son aceptables para las personas pero muy críticas para un sistema de reconocimiento automático del habla.

En la publicación de Kato, M. y otros, "A Wideband Noise Suppressor for the AMR Wideband Speech Codec", Speech coding, 2002, IEEE workshop proceedings, 06.10.02, la ganancia espectral como se define por parte de Ephraim y Malah se modifica con un escalado y una limitación condicional.

Objeto y resumen de la presente invención

La intención de la presente invención es por tanto proporcionar un procedimiento de reducción de ruido para el reconocimiento automático del habla y que, al mismo tiempo, reduce el ruido musical del espectro de potencia del habla limpiada.

Este objetivo se alcanza con la presente invención por el hecho de que se refiere a un procedimiento de reducción de ruido para el reconocimiento automático del habla, como se define en la reivindicación 1, a un sistema de reconocimiento automático del habla, como se define en la reivindicación 12, y a un producto programa de ordenador, como se define en la reivindicación 13.

La presente invención cumple las necesidades arriba mencionadas puesto que utiliza una técnica de atenuación espectral en lugar de una técnica de substracción espectral, eliminando de esta forma el problema del ruido musical, y que la implementación de una regla de atenuación espectral de Ephraim-Malah modificada reduce la distorsión espectral introducida por la regla original en las partes vocales de las señales, obteniendo de esta forma mejores rendimientos cuando se utiliza en un sistema automático de reconocimiento del habla.

Breve descripción de las figuras

Para una mejor comprensión de la presente invención, a continuación se describirá una realización preferida, la cual se pretende puramente a modo de ejemplo y no se debe considerar de forma limitativa, con referencia a las figuras adjuntas, en las cuales:

- la figura 1 muestra un diagrama de bloques de las fuentes comunes de degradación del habla;

- la figura 2 muestra un diagrama de bloques de la reducción de ruido para el reconocimiento automático del habla;

- las figuras 3 y 4 muestran gráficos de un factor de sobreestimación de ruido y de un factor de fondo espectral como función de una relación de señal a ruido global que se utilizan en el procedimiento de reducción de ruido según la presente invención;

- la figura 5 muestra una regla de atenuación espectral de Ephraim-Malah estándar; y

- las figuras 6-10 muestran una regla de atenuación espectral de Ephraim-Malah modificada según la presente invención a diferentes relaciones globales de señal a ruido.

Descripción detallada de realizaciones preferidas de la presente invención

La siguiente discusión se presenta para permitir a una persona experta en la técnica utilizar la presente invención. Para la personas expertas en la técnica se harán fácilmente aparentes varias modificaciones de las realizaciones sin salir del ámbito de la presente invención como se reivindica. Por tanto, no se pretende que la presente invención se limite a las realizaciones que se muestran, sino que se debe contemplarse el ámbito más amplio consistente con los principios y características que aquí se describen y que se definen en las reivindicaciones adjuntas.

La presente invención se refiere a un sistema automático de reconocimiento del habla que comprende un sistema de reducción de ruido basado en la técnica de atenuación espectral, y en concreto en la regla de atenuación de Ephraim-Malah, en el cual la fórmula global de la ganancia $G_k(\gamma_k, \xi_k)$ permanece sin cambios, mientras que las estimaciones de las relaciones de señal a *ruido a priori* y *a posteriori* $\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$ se modifican haciéndolas dependientes de un factor de ponderación de ruido $\alpha(m)$ y de un factor de fondo espectral $\beta(m)$, de la forma siguiente:

$$\hat{\gamma}_k(m) = \max\left(\frac{|Y_k(m)|^2}{\alpha(m)|\hat{D}_k(m)|^2} - 1, \beta(m)\right) + 1 \qquad (9)$$

$$\hat{\xi}_k(m) = \max\left(\eta(m)\frac{|\hat{X}_k(m-1)|^2}{\alpha(m)|\hat{D}_k(m-1)|^2} + (1-\eta(m))[\hat{\gamma}_k(m) - 1], \beta(m)\right) \qquad (10)$$

$$, \eta(m) \in [0,1)$$

donde:

- $|Y_k(m)|^2$ es la línea espectral k-ésima del espectro de potencia del habla ruidosa;

- $|\hat{X}_k(m)|^2$ es la línea espectral k-ésima de la estimación del espectro de potencia del habla limpiada;

- $|\hat{D}_k(m)|^2$ es la línea espectral k-ésima de la estimación del espectro de potencia del ruido aditivo;

- $\hat{\xi}_k(m)$ es la estimación de la relación de señal a ruido *a priori* relativa a la línea espectral k-ésima;

- $\hat{\gamma}_k(m)$ es la estimación de la relación de señal a ruido *a posteriori* relativa a la línea espectral k-ésima;

- $\alpha(m)$ es el factor de peso de ruido para ponderar, concretamente sobreestimar o subestimar, la estimación $|\hat{D}_k(m)|^2$ del espectro de potencia del ruido en el cálculo de las estimaciones $\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$ de las relaciones de señal a ruido *a priori* y *a posteriori*;

- $\beta(m)$ es el factor de fondo espectral para el fondo de las estimaciones $\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$ de las relaciones de señal a ruido *a priori* y *a posteriori*; y

- $\eta(m)$ es un coeficiente de peso para ponderar adecuadamente los dos términos de la fórmula (10).

El factor de ponderación de ruido $\alpha(m)$ y el factor de fondo espectral $\beta(m)$ son funciones de la relación de señal a ruido global SNR(m), que se define como:

$$SNR(m) = 10\log_{10}\left(\frac{\sum_k |Y_k(m)|^2}{\sum_k |\hat{D}_k(m)|^2}\right) \qquad (11)$$

Las figuras 3 y 4 muestran un desarrollo preferido del factor de ponderación de ruido $\alpha(m)$ y del factor de fondo espectral $\beta(m)$ respecto a la relación de señal a ruido global SNR(m). El factor de ponderación de ruido $\alpha(m)$ y el factor de fondo espectral $\beta(m)$ son funciones lineales a tramos y se pueden definir de la siguiente forma:

$$\alpha(m) = \begin{cases} 1{,}5 & \text{si } SNR(m) < 0 \\ 1{,}5 - \frac{(1{,}5 - 0{,}001)}{20} \cdot SNR(m) & \text{si } 0 \leq SNR(m) \leq 20 \\ 0{,}001 & \text{si } SNR(m) > 20 \end{cases} \qquad (12)$$

$$\beta(m) = \begin{cases} 0{,}001 & \text{si } SNR(m) < 0 \\ \dfrac{(1{,}0 - 0{,}01)}{20} \cdot SNR(m) & \text{si } 0 \leq SNR(m) \leq 20 \\ 1{,}0 & \text{si } SNR(m) > 20 \end{cases} \qquad (13)$$

Los valores que se indican en las fórmulas (12) y (13) se indican puramente a modo de ejemplo y no se deben considerar limitativos. En general, se podrían utilizar otros valores de forma útil, mientras se mantenga el desarrollo general del factor de ponderación de ruido α(m) y del factor de fondo espectral β(m) respecto a la relación de señal a ruido global SNR(m).

En concreto, el factor de ponderación de ruido α(m) respecto a la relación de señal a ruido global SNR(m) debería tener un primer valor sustancialmente constante cuando la relación de señal a ruido global SNR(m) es menor que un primer umbral, un segundo valor sustancialmente constante menor que el primer valor sustancialmente constante cuando la relación de señal a ruido global SNR(m) es mayor que un segundo umbral, y valores decrecientes desde el primer valor sustancialmente constante hasta el segundo valor sustancialmente constante cuando la relación de señal a ruido global SNR(m) aumenta desde el primer umbral hasta el segundo umbral.

El factor de fondo espectral β(m) respecto a la relación de señal a ruido global SNR(m) debería tener un primer valor sustancialmente constante cuando la relación de señal a ruido global SNR(m) es menor que un primer umbral, un segundo valor sustancialmente constante mayor que el primer valor sustancialmente constante cuando la relación de señal a ruido global SNR(m) es mayor que un segundo umbral, y valores crecientes desde el primer valor sustancialmente constante hasta el segundo valor sustancialmente constante cuando la relación de señal a ruido global SNR (m) aumenta desde el primer umbral hasta el segundo umbral. Los desarrollos pueden ser por tramos de líneas rectas, como se muestra en las figuras 3 y 4, o pueden ser por líneas curvas similares a las de las figuras 3 y 4, es decir, líneas curvas en las que el tramo intermedio no constante es lineal, como en las figuras 3 y 4, o curvado, por ejemplo una curva de tipo coseno o seno, y se redondean o suavizan las transiciones desde el tramo intermedio no constante hasta los tramos constantes.

La estimación $|\hat{D}_k(m)|^2$ del espectro de potencia del habla ruidosa de las fórmulas (9), (10) y (11) se calcula por medio de una recursión de primer orden como se describe en la publicación anteriormente mencionada “Noise Estimation Techniques for Robust Speech Recognition”.

Preferiblemente, la recursión de primer orden se puede implementar junto con un detector de actividad vocal estándar basado en energía, que es un sistema bien conocido que detecta la presencia o ausencia de habla basándose en una comparación de la energía total de la señal de habla con un umbral adaptativo y genera una bandera booleana (VAD) con un valor de “cierto” cuando se encuentra presente una voz y un valor de “falso” cuando la voz se encuentra ausente. Cuando se utiliza un detector de actividad vocal estándar basado en energía, la estimación $|\hat{D}_k(m)|^2$ del espectro de potencia del habla ruidosa se puede calcular de la siguiente forma:

$$|\hat{D}_k(m)|^2 = \begin{cases} \lambda|\hat{D}_k(m-1)|^2 + (1-\lambda)|Y_k(m)|^2 & \text{si } \left\{\left||Y_k(m)|^2 - |\hat{D}_k(m)|^2\right| \leq \mu\sigma(m)\right\} \\ & \quad \wedge \{VAD = \text{falso}\} \\ |\hat{D}_k(m-1)|^2 & \text{en otro caso} \end{cases} \qquad (14)$$

donde λ es un factor de ponderación que controla la velocidad de actualización de la recursión y vale entre 0 y 1, preferiblemente un valor de 0,9. μ es un factor de multiplicación que controla la dinámica permitida del ruido y presenta preferiblemente un valor de 4,0, y σ(m) es la desviación típica del ruido, que se estima de la siguiente forma:

$$\sigma^2(m) = \lambda\,\sigma^2(m-1) + (1-\lambda)\left(|Y_k(m)|^2 - |\hat{D}_k(m)|^2\right)^2 \qquad (15)$$

La figura 5 muestra la regla de atenuación espectral de Ephraim-Malah estándar (G_k, $\xi_k(m)$ y $\gamma_k(m)$ calculados según las fórmulas (3), (7) y (8)), mientras que las figuras 6-10 muestran la regla de atenuación espectral de Ephraim-Malah modificada según la presente invención (G_k, $\xi_k(m)$ y $\gamma_k(m)$ calculados según las fórmulas (3), (10) y (9)) a diferentes relaciones de señal a ruido globales SNR(m) (0, 5, 10, 15 y 20 dB). Se puede apreciar por parte de la persona experta en la técnica que el efecto de la modificación introducida es una reducción gradual de la atenuación producida por la ganancia original en las zonas en las que la relación de señal a ruido *a posteriori* $\gamma_k(m)$ es alta, puesto que aumenta la relación de señal a ruido SNR(m) global.

Se ha realizado un extenso trabajo experimental para validar la presente invención, y a continuación se relatan algunos resultados, que pueden ser útiles para destacar las características de la invención.

En concreto, los experimentos se llevaron a término con un sistema de reconocimiento automático del habla, utilizando reducción de ruido con la atenuación espectral de Ephraim-Malah estándar y con la reducción de ruido que se propone en la presente invención. El sistema automático de reconocimiento del habla se ha entrenado para los idiomas objetivo utilizando corpus extensos, independientes de dominio y función, no recogidos en ambientes ruidosos y sin ruido añadido.

El experimento se realizó sobre el corpus Aurora3, que es el corpus estándar definido por el ETSI Aurora Project para las pruebas de reducción de ruido, y que consiste en dígitos conectados registrados en un coche en varios idiomas (italiano, español y alemán). Se utilizó un conjunto de prueba de alto error de coincidencia y un componente ruidoso del conjunto de entrenamiento (como conjunto de prueba).

La modificación de la regla de atenuación espectral de Ephraim-Malah según la presente invención produce una media de reducción del error de 28,9% respecto a la substracción espectral de Wiener del estado del arte, y una media de reducción del error de 22,9% respecto a la regla de atenuación espectral de Ephraim-Malah estándar. La media de reducción de error respecto a la no reducción de ruido es 50,2%.

Finalmente, es claro que se pueden realizar numerosas modificaciones y variantes a la presente invención, entrando todas dentro del ámbito de la presente invención, como se define en las reivindicaciones adjuntas.

Referencias citadas en la presente descripción

Esta lista de referencias citadas por el solicitante es solamente para la conveniencia del lector. No forma parte del documento de Patente Europea. Aunque se ha prestado mucha atención en la recopilación de las referencias, no se pueden descartar errores u omisiones y la Oficina Europea de Patentes renuncia a cualquier responsabilidad respecto a la misma.

Documentos de patente citados en la presente descripción

- US 2002/0002455 A [0020]
- WO 0152242 A [0021]

Textos no de patente citados en la presente descripción

- H. **HERMANSKY**; N. **MORGAN**. RASTA Processing of Speech. IEEE Transactions on Speech and Audio Processing, volumen 2, número 4, 1994. [0004]
- N. **VIRAG**. Single Channel Enhancement Based on Masking Properties of the Human Auditory System. IEEE Transactions on Speech and Audio Processing, volumen 7, número 2, 1999. [0007]
- V. **SCHLESS**; F. **CLASS**. SNR-Dependent Flooring and Noise Overestimation for Joint Application of Spectral Substraction and Model Combination. ICSLP, 1998. [0012]
- Y. **EPHRAIM**; D. **MALAH**. Speech Enhancement Using a Minimum Min-Square Error Log-Spectral Amplitude Estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, volumen 33, número 2, páginas 443-445, 1985. [0014]
- H. G. **HIRCH**; C. **EHRLICHER**. Noise Estimation Techniques for Robust Speech Recognition. ICASSP 1995, páginas 153-156. [0016]
- Y. **EPHRAIM**; D. **MALAH**. Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing, volumen 32, número 6, páginas 1109-1121, 1984. [0018]
- **KATO**. M. y otros. A Wideband Noise Suppressor for the AMR Wideband Speech Codec. IEEE workshop proceedings, 06 octubre 2002. [0024]

REIVINDICACIONES

1. Procedimiento de reducción del ruido para reconocimiento automático del habla, que comprende:

- calcular un espectro de magnitud $|Y_k(m)|$ de un habla ruidosa que contiene un habla limpia a ser reconocida y ruido que afecta al habla limpia;

- calcular un espectro de potencia ($|Y_k(m)|^2$) del habla ruidosa;

- calcular una estimación ($|\hat{X}_k(m)|^2$) de un espectro de potencia del habla limpia;

- calcular una estimación ($|\hat{D}_k(m)|^2$) de un espectro de potencia del ruido;

- calcular una estimación ($\hat{\xi}_k(m)$) de una relación de señal a ruido *a priori* como función de la estimación ($|\hat{X}_k(m)|^2$) del espectro de potencia del habla limpia y de la estimación ($|\hat{D}_k(m)|^2$) del espectro de potencia del ruido;

- calcular una estimación ($\hat{\gamma}_k(m)$) de una relación de señal a ruido *a posteriori* como función del espectro de potencia ($|Y_k(m)|^2$) del habla ruidosa y de la estimación ($|\hat{D}_k(m)|^2$) del espectro de potencia del ruido;

- calcular una ganancia de atenuación ($G_k(m)$) como función de la estimación ($\hat{\xi}_k(m)$) de la relación de señal a ruido *a priori* y de la estimación ($\hat{\gamma}_k(m)$) de la relación de señal a ruido *a posteriori*;

- calcular una estimación ($|\hat{X}_k(m)|$) de un espectro de magnitud del habla limpia como función del espectro de magnitud ($|Y_k(m)|$) del habla ruidosa y de la ganancia de atenuación ($G_k(m)$);

caracterizado por el hecho de que calcular las estimaciones ($\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$) de las relaciones de señal a ruido *a priori* y *a posteriori* comprende:

- calcular un factor de ponderación de ruido ($\alpha(m)$) para ponderar la estimación ($|\hat{D}_k(m)|^2$) del espectro de potencia del ruido y calcular las estimaciones ($\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$) de las relaciones de señal a ruido *a priori* y *a posteriori*;

- calcular un factor de fondo espectral ($\beta(m)$) para solar las estimaciones ($\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$) de las relaciones de señal a ruido *a priori* y *a posteriori*; y

- calcular las estimaciones ($\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$) de las relaciones de señal a ruido *a priori* y *a posteriori* también como función del factor de ponderación del ruido ($\alpha(m)$) y el factor de fondo espectral ($\beta(m)$).

2. Procedimiento de reducción de ruido como se describe en la reivindicación 1, en el que el factor de ponderación del ruido ($\alpha(m)$) y el factor de fondo espectral ($\beta(m)$) se calculan como función de la relación de señal a ruido global (SNR(m)).

3. Procedimiento de reducción de ruido como se reivindica en la reivindicación 2, en el que el factor de ponderación del ruido ($\alpha(m)$) respecto a la relación de señal a ruido global (SNR(m)) presenta un primer valor sustancialmente constante cuando la relación de señal a ruido global (SNR(m)) es menor que un primer umbral, un segundo valor sustancialmente constante menor que el primer valor sustancialmente constante cuando la relación de señal a ruido global (SNR(m)) es mayor que un segundo umbral, y valores decrecientes cuando la relación de señal a ruido global (SNR(m)) se encuentra entre el primer y el segundo umbral.

4. Procedimiento de reducción de ruido como se reivindica en la reivindicación 3, en el que el factor de ponderación del ruido ($\alpha(m)$) decrece linealmente cuando la relación de señal a ruido global (SNR(m)) se encuentra entre el primer y el segundo umbral.

5. Procedimiento de reducción de ruido como se reivindica en cualquiera de las reivindicaciones 2 a 4, en el que el factor de fondo espectral ($\beta(m)$) respecto a la relación de señal a ruido global (SNR(m)) presenta un primer valor sustancialmente constante cuando la relación de señal a ruido global (SNR(m)) es menor que un primer umbral, un segundo valor sustancialmente constante mayor que el primer valor sustancialmente constante cuando la relación de señal a ruido global (SNR(m)) es mayor que un segundo umbral, y valores crecientes cuando la relación de señal a ruido global (SNR(m)) se encuentra entre el primer y el segundo umbral.

6. Procedimiento de reducción de ruido como se reivindica en la reivindicación 5, en el que el factor de fondo espectral ($\beta(m)$) crece linealmente cuando la relación de señal a ruido global (SNR(m)) se encuentra entre el primer y el segundo umbral.

7. Procedimiento de reducción de ruido como se reivindica en cualquiera de las reivindicaciones anteriores, en el que la estimación ($\hat{\gamma}_k(m)$) de la relación de señal a ruido *a posteriori* se calcula de la siguiente forma:

$$\hat{\gamma}_k(m) = \max\left(\frac{|Y_k(m)|^2}{\alpha(m)|\hat{D}_k(m)|^2} - 1, \beta(m)\right) + 1$$

donde:

- $\hat{\gamma}_k(m)$ es la estimación de la relación de señal a ruido *a posteriori* de la k-ésima línea espectral;

- $|Y_k(m)|^2$ es la k-ésima línea espectral del espectro de potencia del habla ruidosa;

- $|\hat{D}_k(m)|^2$ es la k-ésima línea espectral de la estimación del espectro de potencia del ruido;

- $\alpha(m)$ es el factor de ponderación del ruido;

- $\beta(m)$ es el factor de fondo espectral;

- k es el índice de las líneas espectrales de los espectros; y

- m es el índice de las ventanas temporales dentro de las cuales se procesa el habla ruidosa para la reducción del ruido.

8. Procedimiento de reducción de ruido como se reivindica en cualquiera de las reivindicaciones anteriores, en el que la estimación ($\hat{\xi}_k(m)$) de la relación de señal a ruido *a priori* se calcula de la siguiente forma:

$$\hat{\xi}_k(m) = \max\left(\eta(m)\frac{|\hat{X}_k(m-1)|^2}{\alpha(m)|\hat{D}_k(m-1)|^2} + (1-\eta(m))[\hat{\gamma}_k(m) - 1], \beta(m)\right)$$

$$, \ \eta(m) \in [0,1)$$

donde:

- $\hat{\xi}_k(m)$ es la estimación de la relación de señal a ruido *a priori* para la k-ésima línea espectral;

- $\hat{\gamma}_k(m)$ es la estimación de la relación de señal a ruido *a posteriori* para la k-ésima línea espectral del espectro de potencia del habla ruidosa;

- $|\hat{X}_k(m)|^2$ es la k-ésima línea espectral de la estimación del espectro de potencia del habla limpia;

- $|\hat{D}_k(m)|^2$ es la k-ésima línea espectral de la estimación del espectro de potencia del ruido;

- $\alpha(m)$ es el factor de ponderación del ruido;

- $\beta(m)$ es el factor de fondo espectral;

- k es el índice de las líneas espectrales de los espectros; y

- m es el índice de las ventanas temporales dentro de las cuales se procesa el habla ruidosa para la reducción del ruido.

9. Procedimiento de reducción de ruido como se reivindica en cualquiera de las reivindicaciones anteriores, en el que la ganancia de atenuación ($G_k(m)$) se calcula de la forma siguiente:

$$G_k(m) = \frac{\hat{\xi}_k(m)}{1 + \hat{\xi}_k(m)} \exp\left(\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right)$$

donde:

- $G_k(m)$ es la ganancia de Ephraim-Malah para la k-ésima línea espectral;

- $\hat{\xi}_k(m)$ es la estimación de la relación de señal a ruido *a priori* para la k-ésima línea espectral;

- $\hat{\gamma}_k(m)$ es la estimación de la relación de señal a ruido *a posteriori* para la k-ésima línea espectral;

$$\nu_k(m) = \frac{\hat{\xi}_k(m)}{1 + \hat{\xi}_k(m)} \hat{\gamma}_k(m)$$

- k es el índice de las líneas espectrales de los espectros; y

- m es el índice de las ventanas temporales dentro de las cuales se procesa el habla ruidosa para la reducción del ruido.

10. Procedimiento de reducción de ruido como se reivindica en cualquiera de las reivindicaciones anteriores, en el que la estimación ($|\hat{D}_k(m)|^2$) del espectro de potencia del ruido se calcula de la forma siguiente:

$$|\hat{D}_k(m)|^2 = \begin{cases} \lambda|\hat{D}_k(m-1)|^2 + (1-\lambda)|Y_k(m)|^2 & \text{si } \left\{\left||Y_k(m)|^2 - |\hat{D}_k(m)|^2\right| \leq \mu\sigma(m)\right\} \\ & \wedge \{\text{VAD} = \text{falso}\} \\ |\hat{D}_k(m-1)|^2 & \text{en otro caso} \end{cases}$$

donde:

- $|\hat{D}_k(m)|^2$ es la k-ésima línea espectral de la estimación del espectro de potencia del ruido;

- $|Y_k(m)|^2$ es la k-ésima línea espectral del espectro de potencia del habla ruidosa;

- λ es un factor de ponderación que controla la velocidad de actualización de la recursión;

- μ es un factor de multiplicación que controla la dinámica permitida del ruido; y

- $\sigma(m)$ es la desviación típica del ruido, la cual se estima de la forma siguiente:

$$\sigma^2(m) = \lambda\,\sigma^2(m-1) + (1-\lambda)\left(|Y_k(m)|^2 - |\hat{D}_k(m)|^2\right)^2$$

11. Procedimiento de reducción de ruido como se reivindica en la reivindicación 2, en el que la relación de señal a ruido global (SNR(m)) se calcula de la forma siguiente:

$$\text{SNR}(m) = 10\log_{10}\left(\frac{\sum_k |Y_k(m)|^2}{\sum_k |\hat{D}_k(m)|^2}\right)$$

donde:

- SNR(m) es la relación de señal a ruido global;

- $|\hat{D}_k(m)|^2$ es la k-ésima línea espectral de la estimación del espectro de potencia del ruido; y

- $|Y_k(m)|^2$ es la k-ésima línea espectral del espectro de potencia del habla ruidosa.

12. Sistema automático de reconocimiento del habla que comprende un sistema de reducción de ruido configurado para implementar el procedimiento según cualquiera de las reivindicaciones anteriores.

13. Producto programa de ordenador que comprende un código de programa de ordenador capaz, cuando se carga en un sistema de procesado, de implementar el procedimiento según cualquiera de las reivindicaciones 1 a 11.

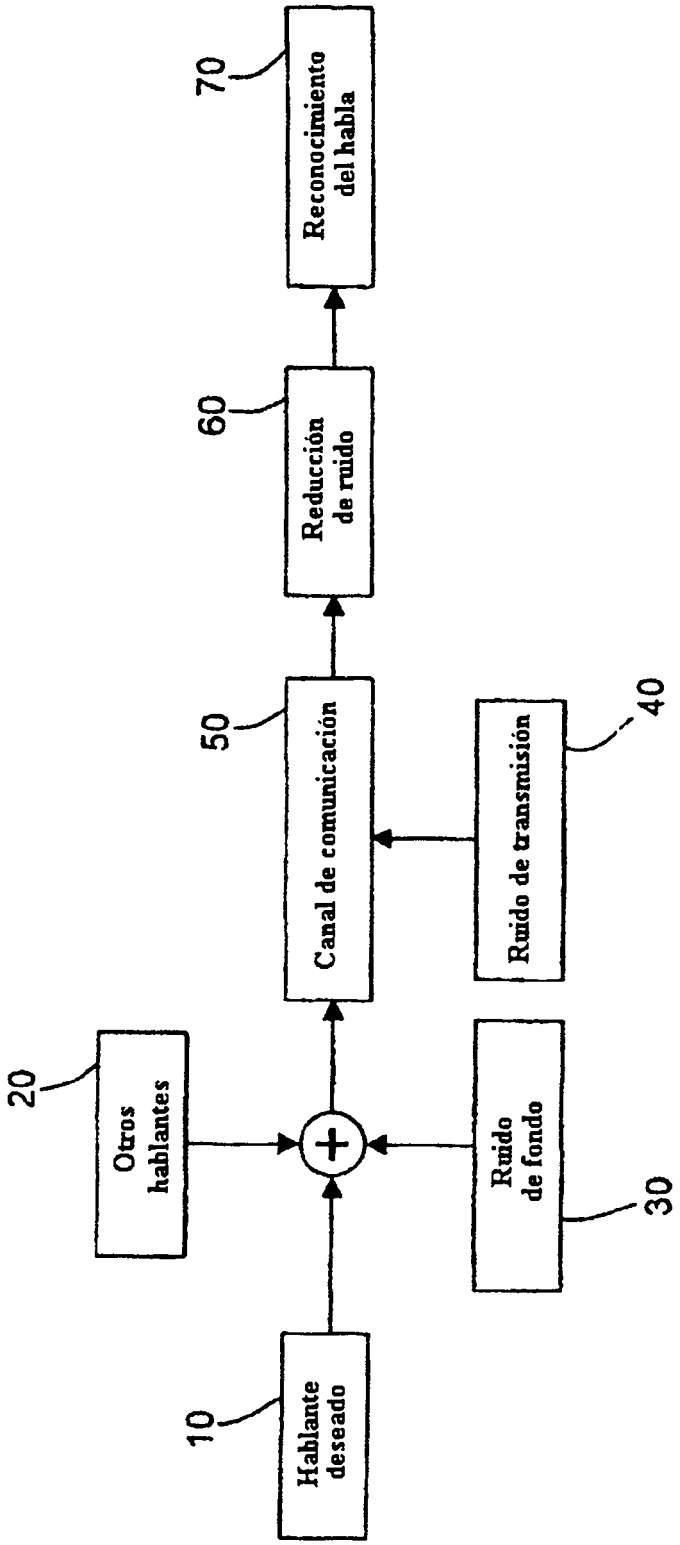
Hablante deseado
10
Otros hablantes
20
Ruido de fondo
30
Ruido de transmisión
40
Canal de comunicación
50
Reducción de ruido
60
Reconocimiento del habla
70

Fig. 1

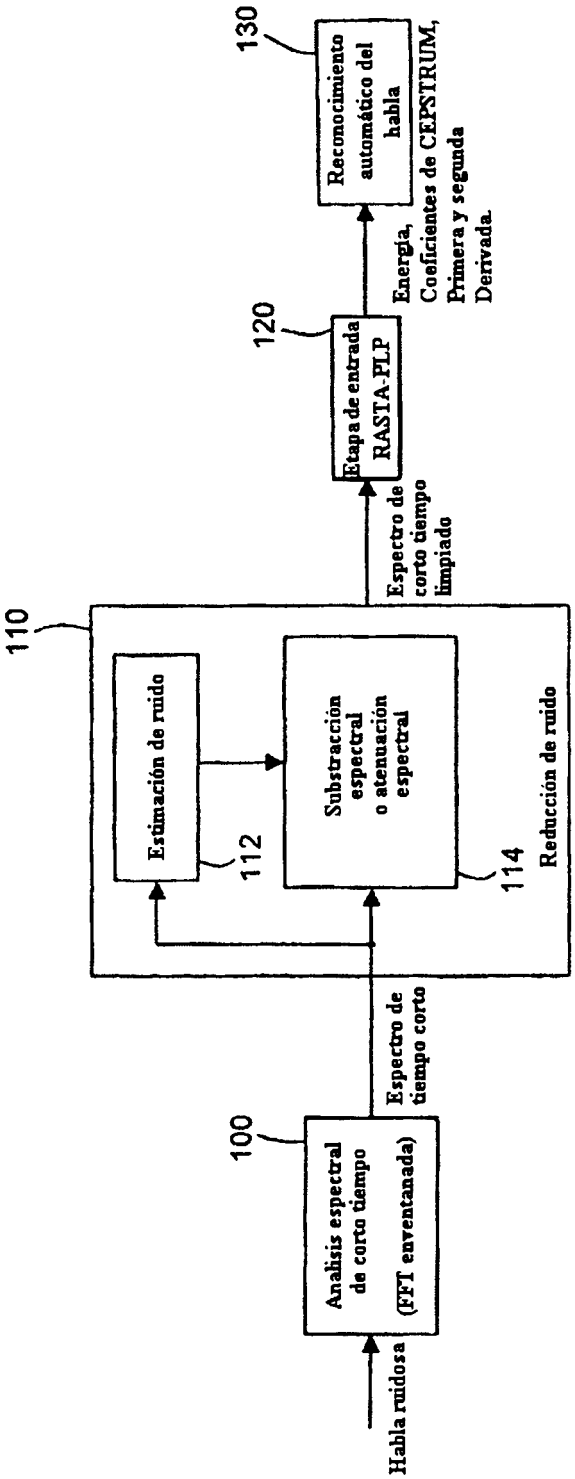
Habla ruidosa
Analisis espectral de corto tiempo (FFT enventanada)
100
Espectro de tiempo corto
110
Estimación de ruido
112
Substracción espectral o atenuación espectral
114
Reducción de ruido
Espectro de corto tiempo limpiado
Etapa de entrada RASTA-PLP
120
Energía, Coeficientes de CEPSTRUM, Primera y segunda Derivada.
Reconocimiento automático del habla
130

Fig. 2

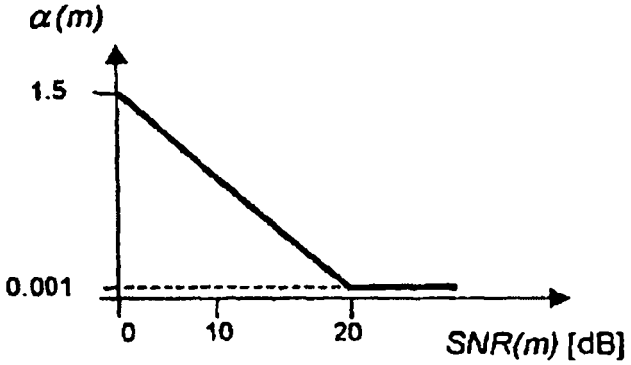

Fig.3

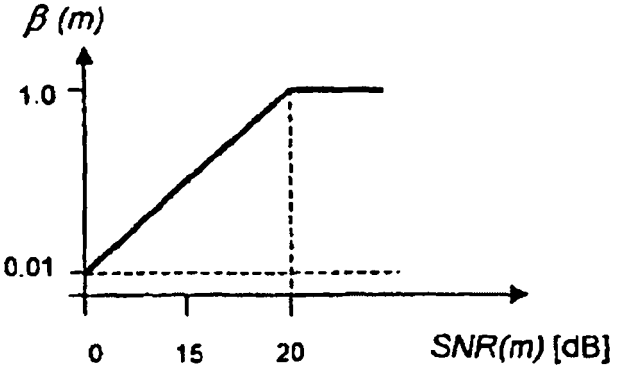

Fig.4

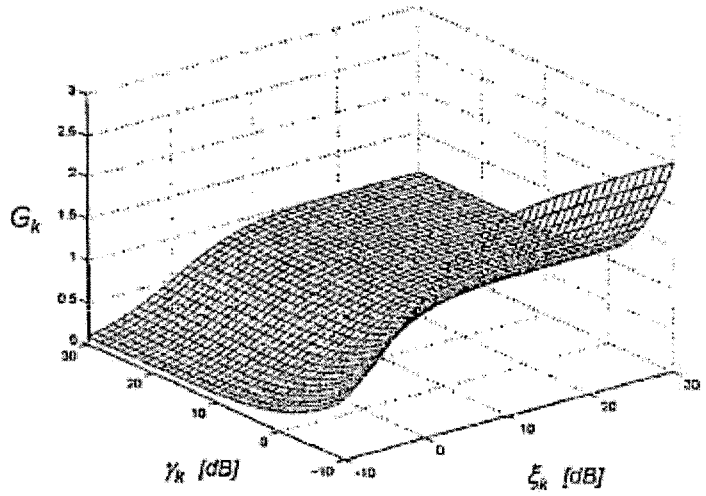

Fig.5

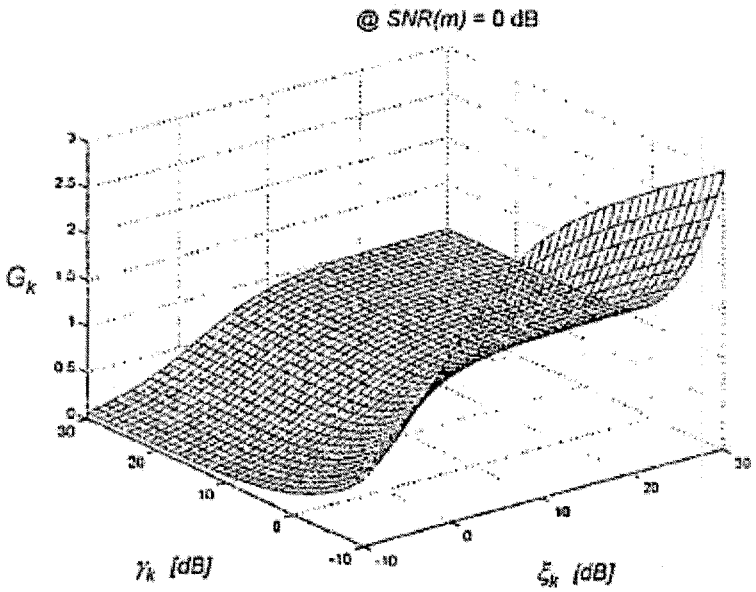

Fig.6

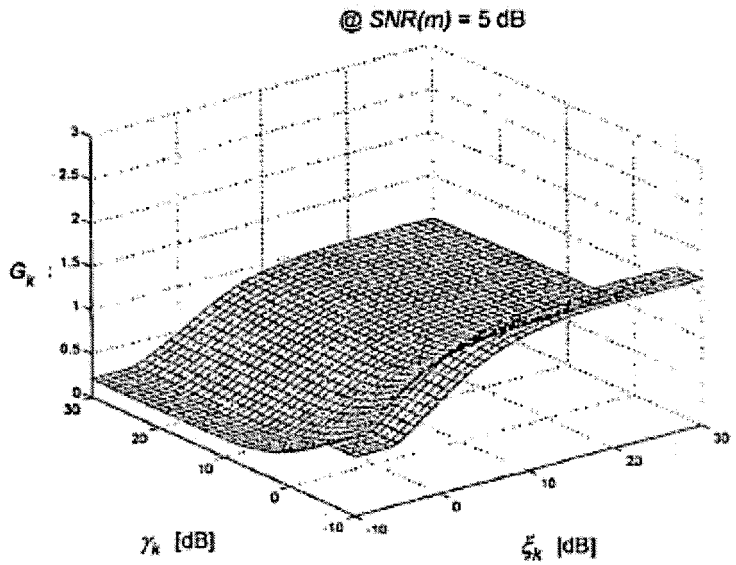

Fig.7

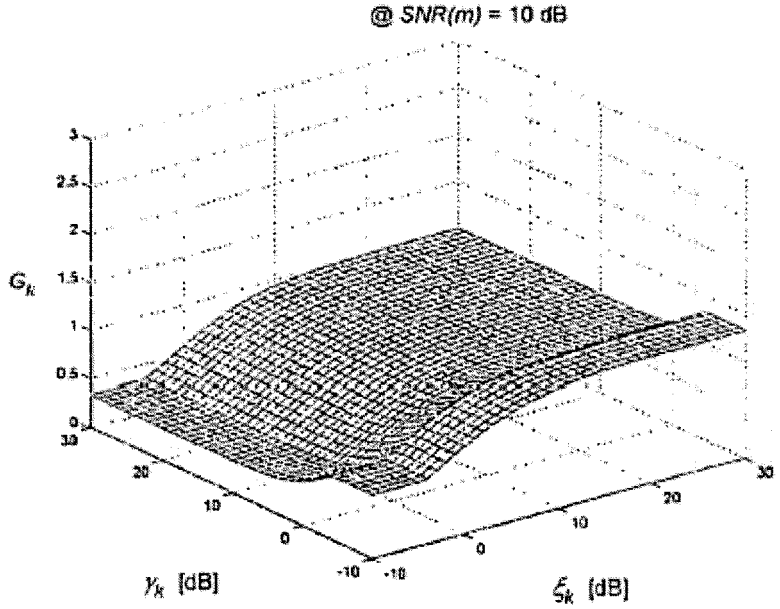

Fig.8

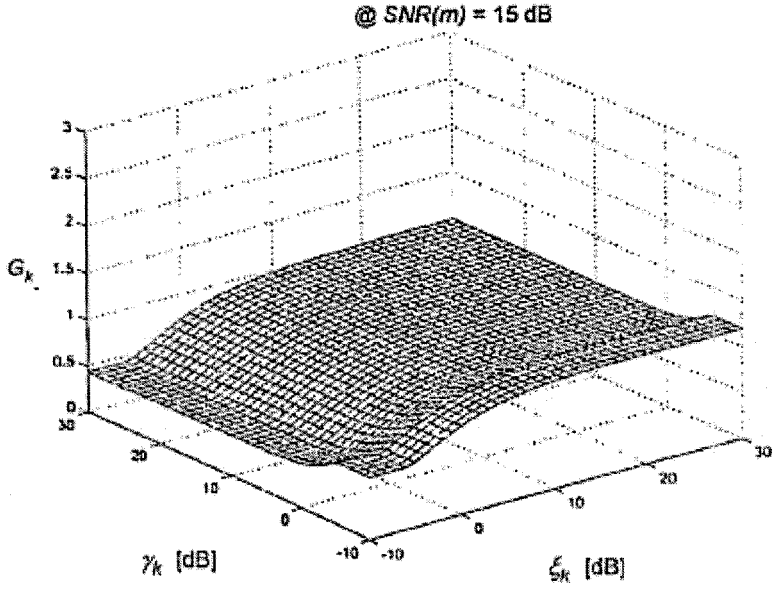

Fig.9

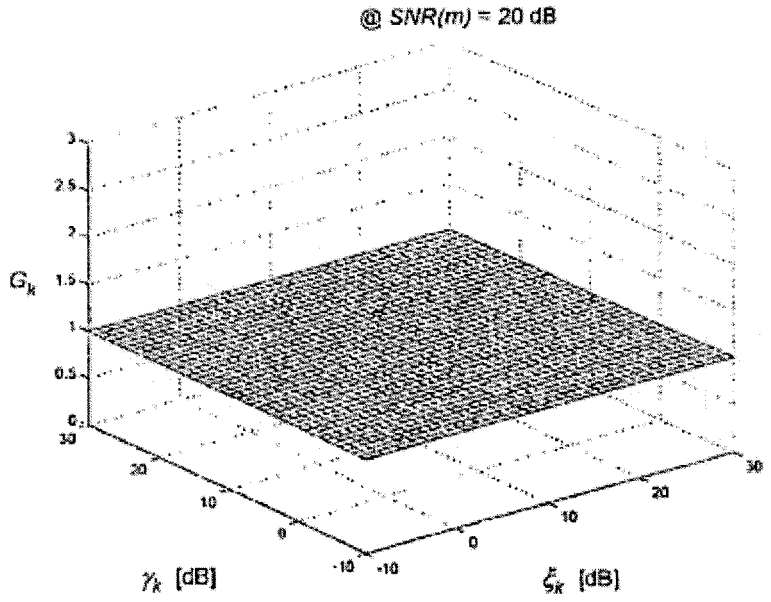

Fig.10