(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 678 415**

(51) Int. CI.:

G10L 21/0208 (2013.01)
G10L 25/30 (2013.01)
G10L 21/0216 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **29.09.2008 E 08017124 (2)**

(97) Fecha y número de publicación de la concesión europea: **25.04.2018 EP 2151822**

(54) Título: **Aparato y procedimiento para procesamiento y señal de audio para mejora de habla mediante el uso de una extracción de característica**

(30) Prioridad:

05.08.2008 US 086361

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
10.08.2018

(73) Titular/es:

FRAUNHOFER-GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E.V. (100.0%)
Hansastrasse 27c
80686 München, DE

(72) Inventor/es:

UHLE, CHRISTIAN;
HELLMUTH, OLIVER;
GRILL, BERNHARD y
RIDDERBUSCH, FALKO

(74) Agente/Representante:

SALVA FERRER, Joan

DESCRIPCIÓN

Aparato y procedimiento para procesamiento y señal de audio para mejora de habla mediante el uso de una extracción de característica

CAMPO DE LA INVENCIÓN

[0001] La presente invención se refiere al campo de procesamiento de señales de audio y, particularmente, al campo de mejora de habla de señales de audio, de tal manera que la señal procesada tenga contenido de habla, que tiene una inteligibilidad de habla objetiva o subjetiva mejorada.

ANTECEDENTES DE LA INVENCIÓN Y TÉCNICA ANTERIOR

[0002] La mejora de habla se emplea en diferentes aplicaciones. Una aplicación prominente es el uso de procesamiento de señales digitales en aparatos de corrección auditiva o prótesis auditivas. El procesamiento de señal digital en las prótesis auditivas ofrece nuevos medios efectivos para la rehabilitación del deterioro de audición. Aparte de calidad de señal acústica superior, las prótesis auditivas digitales permiten la implementación de estrategias de procesamiento de habla específicas. Para muchas de estas estrategias, es conveniente un estimado de la proporción de habla-a-interferencia (SNR = speech-to-noise) del ambiente acústico. Específicamente, se consideran aplicaciones en las que algoritmos complejos para procesamiento de habla se optimizan para ambientes acústicos específicos, pero estos algoritmos pueden fallar en situaciones que no cumplen con las consideraciones específicas. Esto es cierto especialmente para esquemas de reducción de ruido que pueden introducir artefactos de procesamiento en ambientes silenciosos o en situaciones en las que la SNR es inferior a un cierto umbral. Una selección óptima de parámetros de algoritmos de compresión y amplificación puede depender de la proporción de habla-a-interferencia, de tal manera que una adaptación del ajuste de parámetros dependiendo de estimados SNR, ayuda a proporcionar el beneficio. Además, los estimados de SNR pueden emplearse directamente como parámetros de control para esquemas de reducción de ruido, tales como filtrado Wiener o substracción espectral.

[0003] Otras aplicaciones se dan en el campo de mejora de habla del sonido de una película. Se ha encontrado que muchas personas tienen problemas en comprender el contenido de habla de una película, por ejemplo debido a incapacidades o deterioros auditivos. Para seguir la trama de una película, es importante comprender el habla relevante de la pista de audio, por ejemplo monólogos, diálogos, anuncios y narraciones. La gente que tiene dificultad en oír a menudo experimenta esos sonidos de fondo, por ejemplo ruido ambiental y música que están presentes a un nivel muy alto respecto al habla. En este caso, se desea incrementar el nivel de las señales de habla y atenuar los sonidos de fondo o en general, incrementar el nivel de la señal de habla con respecto al nivel total.

[0004] Una estrategia prominente a la mejora de habla es la ponderación espectral, también referida como atenuación espectral a corto plazo, como se ilustra en la Figura 3. La señal de salida y[k] se calcula al atenuar las señales de sub-banda X(ω) de las señales de alimentación x[k] dependiendo de la energía de ruido dentro de las señales de sub-banda.

[0005] A continuación, la señal de alimentación x[k] se considera que es una mezcla aditiva de la señal de habla deseada s[k] y el ruido de fondo b[k].

$$x[k] = s[k] + b[k] \quad . (1)$$

[0006] Mejora de habla es la mejora en la inteligibilidad objetivo y/o calidad subjetiva del habla.

[0007] Una representación de dominio de frecuencia de la señal de alimentación se calcula mediante una Transformada Fourier de Corto Plazo (STFT = Short-term Fourier Transform), otras transformadas de frecuencia-tiempo o un banco de filtros como se indica en 30. La señal de alimentación se filtra a continuación en el dominio de frecuencia según la Ecuación 2, mientras que la respuesta de frecuencia $G(\omega)$ del filtro, se calcula de tal manera que se reduce la energía del ruido. La señal de salida se calcula mediante procesamiento inverso de las transformadas del tiempo-frecuencia o banco de filtros, respectivamente.

$$Y(\omega) = G(\omega)X(\omega) \quad (2)$$

[0008] Pesos espectrales apropiados $G(\omega)$ se calculan en 31 para cada valor espectral mediante el uso del espectro de señal de alimentación $X(\omega)$ y un estimado del espectro de ruido $\hat{B}(\omega)$ o de forma equivalente, mediante el uso de un estimado de la SNR de sub-banda lineal $\hat{R}(\omega)=\hat{S}(\omega)/\hat{B}(\omega)$. El valor espectral ponderado se transforma de nuevo al dominio de tiempo en 32. Ejemplos prominentes de reglas de supresión de ruido son substracción espectral [S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113-120, 1979] y filtrado Wiener. Considerando que la señal de alimentación es una mezcla aditiva de las señales de habla y de ruido y que el habla y ruido no están correlacionados, el valor de ganancia para el procedimiento de substracción espectral se da en la Ecuación 3.

$$G(\omega)=\sqrt{1-\frac{\left|\hat{B}(\omega)\right|^2}{\left|X(\omega)\right|^2}} \qquad (3)$$

[0009] Pesos similares se derivan de estimados de la SNR de sub-banda lineal $\hat{R}(\omega)$ según la Ecuación 4. Canal

$$G(\omega)=\sqrt{\frac{\hat{R}(\omega)}{\hat{R}(\omega)+1}} \qquad (4)$$

[0010] Diversas extensiones a la substracción espectral se han propuesto en el pasado, es decir el uso de un factor de sobre-substracción y parámetro de piso espectral [M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise", Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, 1979], formas generalizadas [J. Lim, A. Oppenheim, "Enhancement and bandwidth compression of noisy speech", Proc. of the IEEE, vol 67, no. 12, pp. 1586-1604, 1979], el uso de criterios perceptuales (por ejemplo, N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. Speech and Audio Proc., vol. 7, no. 2, pp. 126-137, 1999) y substracción espectral de múltiples bandas (por ejemplo S. Kamath, P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise", Proc. of the IEEE Int. Conf. Acoust. Speech Signal Processing, 2002). Sin embargo, la parte crucial de un procedimiento de ponderación espectral es el estimado del espectro de ruido instantáneo o de la SNR de sub-banda, que es tendiente a errores, especialmente si el ruido no es estacionario. Errores de estimación de ruido llevan a ruido residual, distorsiones de los componentes de habla o ruido musical (un artefacto que se ha descrito como "modulado en frecuencia con calidad tonal" [P. Loizou, Speech Enhancement: Theory and Practice, CRC Press, 2007]).

[0011] Una estrategia simple a estimación de ruido es medir y promediar el espectro de ruido durante pausas de habla. Esta estrategia no produce resultados satisfactorios si el espectro de ruido varía con el tiempo durante actividad de habla y si la detección de las pausas del habla falla. Procedimientos para estimar el espectro de ruido incluso durante la actividad de habla se han propuesto en el pasado y pueden clasificarse según P. Loizou, Speech Enhancement: Theory and Practice, CRC Press, 2007 como

- Algoritmos de seguimiento mínimo
- Algoritmos de promedio recursivo en tiempo
- Algoritmos basados en histograma

[0012] La estimación del espectro de ruido mediante el uso de estadísticas mínimas se ha propuesto por R. Martin, "Spectral subtraction based on minimum statistics", Proc. of EUSIPCO, Edingburgh, UK, 1994. El procedimiento se basa en los seguimientos de mínimos locales de la energía de señal en cada sub-banda. Una regla de actualización no lineal para el estimado de ruido y más rápida actualización se ha propuesto por G. Doblinger, "Computationally Efficient Speech Enhancement By Spectral Minima Tracking In Subbands", Proc. of Eurospeech, Madrid, Spain, 1995.

[0013] Algoritmos de promedio recursivos en tiempo estiman y actualizan el espectro de ruido cada vez que la SNR estimada en una banda de frecuencia particular es muy baja. Esto se realiza al calcular de forma recursiva el promedio ponderado del estimado de ruido pasado y el espectro presente. Los pesos se determinan como una función de la probabilidad de que el habla está presente o como una función del SNR estimada en la banda de frecuencia particular, por ejemplo por I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement", IEEE Signal Proc. Letters, vol. 9, no. 1, pp. 12-15, 2002, y por L. Lin, W. Holmes, E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement", Electronic Letters, vol. 39, no. 9, pp. 754-755, 2003.

[0014] Procedimientos basados en histograma se basan en la consideración de que el histograma de la energía de sub-banda a menudo es bimodal. Un modo de baja energía grande acumula valores de energía de segmentos sin habla o con segmentos de baja energía de habla. El modo de alta energía acumula valores de energía de segmentos de habla con voz y ruido. La energía de ruido en una sub-banda particular se determina a partir del modo de baja energía [H. Hirsch, C. Ehrlicher, "Noise estimation techniques for robust speech recognition", Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Detroit, USA, 1995]. Para una revisión creciente extensa se refiere a P. Loizou, Speech Enhancement: Theory and Practice, CRC Press, 2007.

[0015] Procedimientos para estimar SNR de sub-banda con base en aprendizaje supervisado mediante el uso de características de modulación de amplitud, se reportan por J. Tchorz, B. Kollmeier, "SNR Estimation based on amplitude modulation analysis with applications to noise suppression", IEEE Trans. On Speech and Audio Processing, vol. 11, no. 3, pp. 184-192, 2003, y en M. Kleinschmidt, V. Hohmann, "Sub-band SNR estimation using auditory feature processing", Speech Communication: Special Issue on Speech Processing for Hearing Aids, vol. 39, pp. 47-64, 2003.

[0016] Otras estrategias para mejora de habla son el filtrado síncrono de agudos (por ejemplo en R. Frazier, S. Samsam, L. Braida, A. Oppenheim, "Enhancement of speech by adaptive filtering", Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Philadelphia, USA, 1976), el filtrado de Modulación de Espectro Temporal (STM = Spectro Temporal Modulation) (por ejemplo por N. Mesgarani, S. Shamma, "Speech enhancement based on filtering the spectro-temporal modulations", Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Philadelphia, USA, 2005), y el filtrado basado en una representación de modelo sinusoidal de la señal de alimentación (por ejemplo J. Jensen, J. Hansen, "Speech enhancement using a constrained iterative sinusoidal model", IEEE Trans. on Speech and Audio Processing, vol. 9, no. 7, pp. 731-740, 2001).

[0017] Los procedimientos para estimado de la SNR de sub-banda con base en aprendizaje supervisado mediante el uso de características de modulación de amplitud como se notifica en J. Tchorz, B. Kollmeier, "SNR Estimation based on amplitude modulation analysis with applications to noise suppression", IEEE Trans. On Speech and Audio Processing, vol. 11, no. 3, pp. 184-192, 2003, y por M. Kleinschmidt, V. Hohmann, "Sub-band SNR estimation using auditory feature processing", Speech Communication: Special Issue on Speech Processing for Hearing Aids, vol. 39, pp. 47-64, 200312, 13 son desventajosos ya que se requieren dos etapas de procesamiento espectrograma. La primera etapa de procesamiento espectrograma es para generar un espectrograma de tiempo/frecuencia de la señal de audio en dominio de tiempo. A continuación, para generar el espectrograma de modulación, se requiere otra transformada de "tiempo/frecuencia", que transforma la información espectral del dominio espectral en el dominio de modulación. Debido al retraso sistemático inherente y al aspecto de resolución de tiempo/frecuencia inherente a cualquier algoritmo de transformación, esta operación de transformación adicional incurre en problemas.

[0018] Una consecuencia adicional de este procedimiento es que los estimados de ruido son bastante imprecisos en condiciones en las que el ruido no es estacionario y en las que pueden ocurrir diversas señales de ruido.

[0019] La publicación "A Comparison of Composite Feature Under Degraded Speech in Speaker Recognition", J.P. Openshaw et al., proceedings of the International Conference on Acoustics, Speech, and Signal Processing, New York, IEEE, April 27, 1993, pages 371-374 describe una variedad de características y su sensibilidad a los desajustes de ruido entre el modelo y las condiciones de ruido de prueba. La identificación del orador se utiliza para una evaluación del rendimiento, dado que es muy sensible a cambios de característica. Se consideran las características primarias MFCC y PLP, junto con su RASTA y las extensiones de regresión de primer orden.

[0020] La publicación "Sub-band SNR Estimation using Auditory Feature Processing", Michael Kleinschmidt

et al., Speech Communication, vol. 39, No. 1-2, January 1, 2003, pages 47-63, XP055077312 describe un concepto para la estimación de habla a largo plazo a índice de ruido (SNR) en bandas de frecuencia individual que se basa en procedimientos conocidos a partir del reconocimiento de habla automático (ASR). Un modelo de percepción auditiva se utiliza como parte delantera. Además, las celdas sigma-pi motivadas fisiológicamente y de forma psicoacústica se utilizan como características secundarias y una red neural lineal o no-lineal se utiliza como clasificador. La estrategia está basada puramente en modulaciones espectro-temporales lentas. Con este fin, se genera una representación espectro-temporal, que se basa en un modelo de percepción. Se calcula un espectrograma de modulación, valores de media de energía de 10 ms se derivan de un banco de filtro de tono gamma como una extracción de característica de referencia. Las celdas sigma-pi se utilizan como características secundarias basadas en espectrogramas de banda fundamentales para el reconocimiento de palabra aislado.

[0021] Es un objetivo de la presente invención proporcionar un concepto mejorado para mejora de habla.

[0022] Este objetivo se logra por un aparato para procesar una señal de audio según la reivindicación 1, un procedimiento de procesamiento de una señal de audio según la reivindicación 9, un aparato para mejora de habla de una señal de audio según la reivindicación 10, un procedimiento de mejora de habla según la reivindicación 12, un aparato para entrenar un combinador de característica según la reivindicación 13, un procedimiento de formación de un combinador de característica según la reivindicación 14 o un programa informático según la reivindicación 15.

[0023] Según un primer aspecto, este objetivo se consigue por medio de un aparato para el procesamiento de una señal de audio para obtener información de control por sub-banda para un filtro de mejora de habla según la reivindicación 1. Según un segundo aspecto, este objetivo se consigue por medio de un procedimiento de procesamiento de una señal de audio para obtener información de control por sub-banda para un filtro de mejora de habla según la reivindicación 9. Según un tercer aspecto, este objetivo se consigue por un aparato para la mejora de habla en una señal de audio según la reivindicación 10. Según un cuarto aspecto, este objetivo se consigue por un procedimiento de mejora de habla en una señal de audio según la reivindicación 12. Según un quinto aspecto, este objetivo se consigue por medio de un aparato para entrenar un combinador de característica para la determinación de parámetros de combinación del combinador de característica según la reivindicación 13. Según un sexto aspecto, este objetivo se consigue por un procedimiento de formación de un combinador de característica para la determinación de parámetros de combinación del combinador de característica según la reivindicación 14. Según un séptimo aspecto, este objetivo se consigue por medio de un programa informático para llevar a cabo, cuando se ejecuta en un equipo, cualquiera de los procedimientos inventivos.

[0024] La presente invención se basa en el hallazgo de que una información a modo de banda en la forma espectral de la señal de audio dentro de la banda específica, es un parámetro muy útil para determinar información de control para un filtro de mejora de habla. Específicamente, una característica de información de forma espectral determinada a modo de banda para una pluralidad de bandas y para una pluralidad de representaciones espectrales de corto-tiempo subsecuentes, proporciona una descripción característica útil de una señal de audio para procesamiento de mejora de habla de la señal de audio. Específicamente, un conjunto de características de forma espectral, en donde cada característica de forma espectral se asocia con una banda de una pluralidad de bandas espectrales, tales como bandas Bark, o en general, bandas que tienen un ancho de banda variable sobre el intervalo de frecuencia, ya proporciona una característica útil establecida para determinar proporciones de señal/interferencia o ruido para cada banda. Para este objetivo, las características de forma espectral para una pluralidad de bandas se procesan por un combinador de características para combinar estas características mediante el uso de parámetros de combinación para obtener la información de control para el filtro de mejora de habla para una porción de tiempo de la señal de audio para cada banda. De preferencia, el combinador de características incluye una red neural, que se controla por muchos parámetros de combinación, en donde estos parámetros de combinación se determinan en una fase de entrenamiento, que se realiza antes de efectuar actualmente el filtrado con mejora de habla. Específicamente, la red neural realiza un procedimiento de regresión de red neural. Una ventaja específica es que los parámetros de combinación pueden determinarse dentro de una fase de entrenamiento utilizando material de audio, que puede ser diferente del material de audio mejorado en habla actual, de tal manera que la fase de entrenamiento tiene que realizarse solo una sola vez, y después de esta fase de entrenamiento, los parámetros de combinación se ajustan de manera fija y pueden aplicarse a cada señal de audio desconocida que tiene habla, que es comparable con una característica de habla de la señal de entrenamiento. Esta característica de habla por ejemplo, puede ser un lenguaje o un grupo de lenguajes, tales como lenguajes europeos contra lenguajes asiáticos, etc.

[0025] De preferencia, el concepto inventivo estima el ruido al aprender las características del habla mediante el uso de extracción de características y redes neurales, en donde las características extraídas de la invención son

características espectrales de bajo nivel directas, que pueden ser extraídas de una forma eficiente y fácil y de manera importante, pueden ser extraídas sin un retardo inherente de sistema grande, de tal manera que el concepto inventivo es específicamente útil para proporcionar un estimado de ruido preciso o SNR, incluso en una situación en donde el ruido no es estacionario y en donde ocurren diversas señales de ruido.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[0026] Realizaciones preferidas de la presente invención se discuten posteriormente con más detalle en referencia a los dibujos anexos en los que:

La Figura 1 es un diagrama de bloques de un aparato o procedimiento preferido para procesar una señal de audio;
La Figura 2 es un diagrama de bloques de un aparato o procedimiento para entrenar un combinador de características según una realización preferida de la presente invención;
La Figura 3 es un diagrama de bloques para ilustrar un aparato y procedimiento para mejora de habla según una realización preferida de la presente invención;
La Figura 4 ilustra una vista general o panorama del procedimiento para entrenar un combinador de características y para aplicar una regresión de red neural mediante el uso de los parámetros de combinación optimizados;
La Figura 5 es un trazo que ilustra el factor de ganancia como una función del SNR, en donde las ganancias aplicadas (línea sólida) se comparan con las ganancias de substracción espectral (línea punteada) y el filtro Wiener (línea con rayas);
La Figura 6 es una visión general sobre las características por banda de frecuencia y características adicionales preferidas para todo el ancho de banda;
La Figura 7 es un diagrama de flujo para ilustrar una implementación preferida del extractor de características;
La Figura 8 ilustra un diagrama de flujo para ilustrar una implementación preferida del cálculo de los factores de ganancia por valor de frecuencia y el cálculo subsecuente de la porción de señal de audio mejorado en habla;
La Figura 9 ilustra un ejemplo de la ponderación espectral, en donde la señal de alimentación en tiempo, la SNR de sub-banda estimada, la SNR estimada en depósitos de frecuencia después de interpolación, los pesos espectrales y la señal de tiempo procesada, se ilustran; y
La Figura 10 es un diagrama de bloques esquemático de una implementación preferida del combinador de características mediante el uso de una red neural de múltiples capas.

DESCRIPCIÓN DETALLADA DE REALIZACIONES PREFERIDAS

[0027] La Figura 1 ilustra un aparato preferido para procesar una señal de audio 10 para obtener información de control 11 para un filtro de mejora de habla 12. El filtro de mejora de habla puede implementarse de muchas formas, tales como un filtro controlable para filtrar la señal de audio 10, mediante el uso de la información de control por banda de frecuencia para cada una de la pluralidad de bandas de frecuencia, para obtener una señal de salida de audio mejorada de habla 13. Como se ilustra posteriormente, el filtro controlable también puede ser implementado como una conversión de tiempo/frecuencia, en donde factores de ganancia individualmente calculados se aplican a los valores espectrales o bandas espectrales, seguido por una conversión de frecuencia/tiempo subsecuentemente realizada.

[0028] El aparato de la Figura 1 comprende un extractor de características 14 para obtener una secuencia de tiempo de representaciones espectrales de corto-tiempo de la señal de audio y para extraer al menos una característica en cada banda de frecuencia de una pluralidad de bandas de frecuencia para una pluralidad de representaciones espectrales de corto-tiempo, en donde al menos una característica representa una forma espectral de una representación espectral de corto-tiempo en una banda de frecuencia de la pluralidad de bandas de frecuencia. Adicionalmente, el extractor de características 14 puede implementarse para extraer otras características aparte de las características de forma espectral. A la salida del extractor de características 14 existen varias características por espectro de corto-tiempo de audio en donde estas varias características al menos incluyen una característica de forma espectral para cada banda de frecuencia de una pluralidad de al menos 10 o de preferencia más, tales como 20 a 30 bandas de frecuencia. Estas características pueden emplearse como están, o pueden procesarse mediante el uso de un procesamiento promedio o cualquier otro procesamiento, tal como promedio geométrico o promedio aritmético o procesamiento de mediana u otro procesamiento de momentos estadísticos (tales como variancia, asimetría...) para obtener, por cada banda, una característica en bruto de una característica promediada, de tal manera que todas estas características en bruto y/o promediadas se alimentan a un combinador de características 15. El combinador de características 15 combina la pluralidad de características de forma espectral y de preferencia, características adicionales mediante el uso de parámetros de combinación, que pueden proporcionarse mediante una alimentación de parámetros de combinación 16, o que puede ser programado fijo

dentro del combinador de características 15, de tal manera que no se requiera la alimentación de parámetro de combinación 16. A la salida del combinador de características, la información de control para el filtro de mejora de habla para cada banda de frecuencia o "sub-banda" de la pluralidad de bandas de frecuencia o la pluralidad de sub-bandas, se obtiene por una porción de tiempo de la señal de audio.

[0029] De preferencia, el combinador de características 15 se implementa como un circuito de regresión de red neural, pero el combinador de características también puede implementarse como cualquier otro combinador de características controlado de forma estadística o numérica, que aplica cualquier operación de combinación a la salida de características por el extractor de características 14, de tal manera que en el extremo, la información de control requerida, tal que resulta un valor SNR a modo de banda o un factor de ganancia a modo de banda. En la realización preferida de una aplicación de red neural, se requiere una fase de entrenamiento ("fase de entrenamiento" significa una fase en la que se realiza aprendizaje de los ejemplos). En esta fase de entrenamiento, se utiliza un aparato para entrenar un combinador de características 15 como se indica en la Figura 2. Específicamente, la Figura 2 ilustra este aparato para entrenar un combinador de características 15, para determinar parámetros de combinación del combinador de características. Para este fin, el aparato en la Figura 2 comprende el extractor de características 14, que de preferencia se implementa en la misma forma que el extractor de características 14 de la Figura 1. Además, el combinador de características 15 también se implementa de la misma forma que el combinador de características 15 de la Figura 1.

[0030] Además de la Figura 1, el aparato en la Figura 2 comprende un controlador de optimización 20, que recibe como una alimentación, información de control para una señal de audio de entrenamiento como se indica en 21. La fase de entrenamiento se realiza con base en señales de audio de entrenamiento conocidas, que tienen una proporción de habla/ruido o interferencia conocida en cada banda. La porción de habla y la porción de ruido son – por ejemplo – suministradas por separado entre sí y la SNR actual por banda se miden al vuelo, es decir durante la operación de aprendizaje. Específicamente, el controlador de optimización 20 es operativo para controlar el combinador de características, de tal manera que el combinador de características se alimenta con las características del extractor de características 14. Con base en estas características y parámetros de combinación intermedios que vienen de una corrida de iteración anterior, el combinador de características 15 calcula entonces información de control 11. Esta información de control 11 se envía al controlador de optimización y en el controlador de optimización 20 se compara con la información de control 21 para la señal de audio de entrenamiento. Los parámetros de combinación intermedios se varían en respuesta a una instrucción del controlador de optimización 20 y mediante el uso de estos parámetros de combinación variados, se calcula un conjunto adicional de información de control por el combinador de características 15. Cuando la información de control adicional coincide mejor con la información de control para la señal de audio de entrenamiento 21, el controlador de optimización 20 actualiza los parámetros de combinación y envía estos parámetros de combinación actualizados 16 al combinador de características que se van a utilizar en la siguiente corrida como parámetros de combinación intermedios. De forma alternativa, o adicionalmente, los parámetros de combinación actualizados pueden ser almacenados en una memoria para uso adicional.

[0031] La Figura 4 ilustra una vista general o panorama de un procesamiento de ponderación espectral mediante el uso de la extracción de características en el procedimiento de regresión de red neural. Los parámetros w de la red neural se calculan mediante el uso de los valores SNR de sub-banda de referencia R_t y características de los ítems de entrenamiento $x_t[k]$ durante la fase de entrenamiento, que se indica en el lado a mano izquierda de la Figura 4. El estimado de ruido y el filtrado de mejora de habla se muestran en el lado a mano derecha de la Figura 4.

[0032] El concepto propuesto sigue la estrategia de ponderación espectral y utiliza un procedimiento novedoso para el cálculo de los pesos o ponderaciones espectrales. El estimado de ruido o interferencia se basa en un procedimiento de aprendizaje supervisado y utiliza un conjunto de características de la invención. Las características se dirigen a la discriminación de componentes de señal tonal contra ruidosos. Adicionalmente, las características propuestas toman en cuenta la evolución de propiedades de señal en una escala de tiempo más grande.

[0033] El procedimiento de estimación de ruido presentado aquí es capaz de tratar con una variedad de sonidos de fondo no estacionarios. Una estimación SNR robusta en ruido de fondo no estacionario se obtiene mediante extracción de características y un procedimiento de regresión de red neural como se ilustra en la Figura 4. Los pesos de valor real se calculan a partir de estimados de SNR en las bandas de frecuencia cuyo espaciamiento se aproxima a la escala Bark. La resolución espectral de la estimación SNR más bien es gruesa o tosca para permitir la medición de una forma espectral en una banda.

[0034] El lado izquierdo de la Figura 4 corresponde a una fase de entrenamiento que, básicamente tiene que realizarse solo una vez. El procedimiento en el lado a mano izquierda de la Figura 4 indicado como entrenamiento 41, incluye un bloque de cómputo SNR de referencia 21, que genera la información de control 21 para una alimentación de señal de audio de entrenamiento en el controlador de optimización 20 de la Figura 2. El dispositivo de extracción de características 14 en la Figura 4 en el lado de entrenamiento corresponde al extractor de características 14 de la Figura 2. En particular, la Figura 2 se ha ilustrado para recibir una señal de audio de entrenamiento, que consiste en una porción de habla y una porción de fondo. A fin de poder realizar una referencia útil, la porción de fondo b_t y la porción de habla s_t están disponibles por separado entre sí y se agregan mediante un sumador 43 antes de alimentarse al dispositivo de extracción de características 14. De esta manera, la salida del sumador 43 corresponde a la alimentación de la señal de audio de entrenamiento en el extractor de características 14 en la Figura 2.

[0035] El dispositivo de entrenamiento de red neural indicado en 15, 20 corresponde a los bloques 15 y 20 y la conexión correspondiente como se indica en la Figura 2 o como se implementa mediante otras conexiones similares, resulta en un conjunto de parámetros de combinación W, que puede almacenarse en la memoria 40. Estos parámetros de combinación se emplean a continuación en el dispositivo de regresión de red neural 15 que corresponden al combinador de características 15 de la Figura 1 cuando se aplica el concepto inventivo como se indica por la aplicación 42 en la Figura 4. El dispositivo de ponderación espectral en la Figura 4 corresponde al filtro controlable 12 de la Figura 1 y el extractor de características 14 en la Figura 4, el lado a mano derecha corresponde al extractor de características 14 en la Figura 1.

[0036] A continuación se discutirá en detalle una breve realización del concepto propuesto. El dispositivo de extracción de características 14 en la Figura 4 opera del siguiente modo.

[0037] Un conjunto de 21 características diferentes se ha investigado a fin de identificar el mejor conjunto de características para el estimado de la sub-banda SNR. Estas características se combinaron en diversas configuraciones y se evaluaron mediante mediciones objetivo y audición informal. El proceso de selección de características resulta en un conjunto de características que comprenden la energía espectral, el flujo espectral, la planicidad espectral, la asimetría espectral, LPC y coeficientes RASTA-PLP. Las características de energía espectral, flujo, planicidad y asimetría, se calculan a partir del coeficiente espectral correspondiente a la escala de banda crítica.

[0038] Las características se detallan con respecto a la Figura 6. Características adicionales y la característica delta de la energía espectral y la característica delta-delta de la energía espectral filtrada de paso bajo y el flujo espectral.

[0039] La estructura de la red neural empleada en los bloques 15, 20 ó 15 en la Figuras 4, o de preferencia empleada en el combinado de características 15 en la Figura 1 o Figura 2, se discuten en conexión con la Figura 10. En particular, la red neural preferida incluye una capa de neuronas de alimentación 100. En general, n neuronas de alimentación pueden emplearse, es decir una neurona por cada característica de alimentación. De preferencia, la red de neuronas tiene 220 neuronas de alimentación que corresponden al número de características. La red neural comprende además una capa oculta 102 con p neuronas de capa ocultas. En general, p es menor que n y en la realización preferida, la capa oculta tiene 50 neuronas. En el lado de salida, la red neural incluye una capa de salida 104 con q neuronas de salida. En particular, el número de neuronas de salida es igual al número de bandas de frecuencia, de tal manera que cada neurona de salida proporciona una información de control por cada banda de frecuencia tal como información SNR (Proporción de Habla-a-Interferencia) por cada banda de frecuencia. Si, por ejemplo, existen de preferencia 25 bandas de frecuencia diferentes que tienen un ancho de banda, que aumenta de frecuencias bajas a altas, entonces el número de neuronas de salida q será igual a 25. De esta manera, la red neural se aplica para el estimado de la SNR de sub-banda de las características de bajo nivel computadas. La red neural tiene, como se ha establecido anteriormente, 220 neuronas de alimentación y una capa oculta 102 con 50 neuronas. El número de neuronas de salida es igual al número de bandas de frecuencia. De preferencia, las neuronas ocultas incluyen una función de activación, que es la tangente hiperbólica y la función de activación de las neuronas de salida es la identidad.

[0040] En general, cada neurona de la capa 102 ó 104 recibe todas las alimentaciones correspondientes, que son, con respecto a la capa 102, las salidas de todas las neuronas de entrada o de alimentación. Entonces, cada neurona de la capa 102 ó 104 realiza una adición ponderada en donde los parámetros de ponderación corresponden a los parámetros de combinación. La capa oculta puede comprender valores de desviación además de los parámetros. Entonces, los valores de desviación también pertenecen a los parámetros de combinación. En

particular, cada alimentación se pondera por su correspondiente parámetro de combinación y la salida de la operación de ponderación, que se indica por una caja ejemplar 106 en la Figura 10 se alimenta en un sumador 108 dentro de cada neurona. La salida del sumador o una alimentación en una neurona puede comprender una función no lineal 110, que puede colocarse en la salida y/o entrada de una neurona, por ejemplo en la capa oculta según pueda ser el caso.

[0041] Los pesos de la red neural se entrenan en mezclas de señales de habla limpias y ruidos de fondo cuya SNR de referencia se computan utilizando señales separadas. El proceso de entrenamiento se ilustra en el lado a mano izquierda de la Figura 4. Habla y ruido se mezclan con una SNR de 3 dB por ítem y alimentan a la extracción de características. Esta SNR es constante sobre el tiempo y un valor SNR de banda amplia. El conjunto de datos comprende 2.304 combinaciones de 48 señales de habla y 48 señales de ruido de 2,5 segundos de longitud cada una. Las señales de habla se originan de diferentes bocinas con 7 idiomas. Las señales de habla son registros de ruido de tráfico, ruido de la multitud y diversas atmósferas naturales.

[0042] Para una regla de ponderación espectral determinada, son apropiadas dos definiciones de la salida de la red neural: la red neural puede entrenarse utilizando los valores de referencia para la SNR de sub-banda variante en tiempo $R(\omega)$ o con los pesos espectrales $G(\omega)$ (derivados de los valores SNR). Simulaciones con SNR de sub-banda como valores de referencia producen mejores resultados objetivo y mejores calificaciones en audición informal en comparación con redes que se entrenaron con ponderaciones espectrales. La red neural se entrena utilizando 100 ciclos de iteración. Un algoritmo de entrenamiento se emplea en este trabajo, que se basa en gradientes conjugados ajustados en escala.

[0043] Posteriormente se discutirán realizaciones preferidas de la operación de ponderación espectral 12.

[0044] Los estimados SNR de sub-banda son interpolados linealmente con la resolución de frecuencia de los espectros de alimentación y transformados en proporciones lineales $\hat{R}$. La SNR de sub-banda se alisa sobre el tiempo y sobre la frecuencia mediante el uso de filtrado de paso bajo IIR para reducir artefactos, que pueden resultar de errores de estimación. El filtrado de paso bajo sobre la frecuencia se requiere además para reducir el efecto de convolución circular, que ocurre si la respuesta de impulso de la ponderación espectral excede la longitud de los cuadros DFT. Se realiza dos veces, mientras que el segundo filtrado se efectúa en orden inverso (partiendo con la última muestra) de tal manera que el filtro resultante tiene cero fases.

[0045] La Figura 5 ilustra el factor de ganancia como una función de la SNR. La ganancia aplicada (línea sólida) se compara con las ganancias de substracción espectral (líneas punteadas) y el filtro Wiener (línea con rayas).

[0046] Los pesos espectrales se calculan según la regla de substracción espectral modificada en la Ecuación 5 y limitado a -18 dB.

$$G(\omega)=\begin{cases}\dfrac{\hat{R}(\omega)^{a}}{\hat{R}(\omega)^{a}+1} & \hat{R}(\omega)\leq 1\\ \dfrac{\hat{R}(\omega)^{\beta}}{\hat{R}(\omega)^{\beta}+1} & \hat{R}(\omega)>1\end{cases} \qquad (5)$$

[0047] Los parámetros α = 3.5 y β = 1 se determinan experimentalmente. Esta atenuación particular sobre 0 dB SNR se elige para evitar distorsiones de la señal de habla a expensas del ruido residual. La curva de atenuación como una función de SNR se ilustra en la Figura 5.

[0048] La Figura 9 muestra un ejemplo para las señales de alimentación y salida, la SNR de sub-banda estimada y los pesos espectrales.

[0049] Específicamente, la Figura 9 tiene un ejemplo de ponderación espectral: señal de tiempo de alimentación, SNR de sub-banda estimada, SNR estimada en depósitos de frecuencia después de interpolación, pesos espectrales y señal de tiempo procesado.

[0050] La Figura 6 ilustra un panorama sobre las características preferidas que se van a extraer por el extractor de características 14. El extractor de características prefiere para cada baja resolución, una banda de frecuencia, es decir por cada una de las 25 bandas de frecuencia para las cuales se requiere una SNR o valor de ganancia, una característica que representa la forma espectral de la representación espectral de corto-tiempo en la banda de frecuencia. La forma espectral en la banda representa la distribución de energía dentro de la banda y puede ser implementada mediante varias reglas de cálculo diferentes.

[0051] Una característica de forma espectral preferida es la medida de planicidad espectral (SFM = spectral flatness measure), que es el promedio geométrico de los valores espectrales divididos por el promedio aritmético de los valores espectrales. En la definición de promedio geométrico/promedio aritmético, una potencia puede aplicarse a cada valor espectral en la banda antes de realizar la operación de raíz n-ésima o la operación de promediado.

[0052] En general, una medida de planicidad espectral también puede calcularse cuando la potencia para procesar cada valor espectral en la fórmula de cálculo para SFM en el denominador, es superior a la potencia empleada para el numerador. Entonces, ambos el denominador y el numerador pueden incluir una fórmula de cálculo de valor aritmético. De forma ejemplar, la potencia en el numerador es 2 y la potencia en el denominador es 1. En general, la potencia empleada en el numerador solo tiene que ser más grande que la potencia empleada en el denominador para obtener una medida de planicidad espectral generalizada.

[0053] Está claro a partir de este cálculo que SFM para una banda en la que la energía se distribuye igualmente sobre toda la banda de frecuencia es menor que 1 y para muchas líneas de frecuencia, se aproxima a valores pequeños cercanos a 0, mientras que en el caso en el que la energía se concentra en un solo valor espectral dentro de una banda, por ejemplo el valor SFM es igual a 1. De esta manera, un alto valor SFM indica una banda en la que la energía se concentra en una cierta posición dentro de la banda, mientras que un pequeño valor SFM indica que la energía se distribuye igualmente dentro de la banda.

[0054] Otras características de forma espectral incluyen la asimetría espectral, que mide la asimetría de la distribución alrededor de su centroide. Existen otras características que se relacionan con la forma espectral de una representación de frecuencia de corto-tiempo dentro de una cierta banda de frecuencia.

[0055] Mientras que la forma espectral se calcula para una banda de frecuencia, existen otras características, que se calculan para una banda de frecuencia así como se indica en la Figura 6 y como se discute en detalle a continuación. Y, también existen características adicionales, que no necesariamente tienen que calcularse para una banda de frecuencia, pero que se calculan para todo el ancho de banda.

Energía Espectral

[0056] La energía espectral se calcula para cada cuadro de tiempo y banda de frecuencia y se normaliza por energía total del cuadro. Adicionalmente, la energía espectral es filtrada de paso bajo con el tiempo mediante el uso de un filtro IIR de segundo orden.

Flujo espectral

[0057] El flujo espectral *SF* se define como la diferencia entre espectros de cuadros sucesivos 20 y frecuentemente se implementa mediante una función de distancia. En este trabajo, el flujo espectral se calcula mediante el uso de la distancia euclidiana según la Ecuación 6, con coeficientes espectrales $X(m,k)$, índice de marco de tiempo *m*, índice de sub-banda *r*, fronteras inferior y superior de la banda de frecuencia l_r y u_r, respectivamente.

$$SF(m,r) = \sqrt{\sum_{q=l_r}^{u_r} (|X(m,q)| - |X(m-1,q)|)^2} \qquad (6)$$

Medida de planicidad espectral

[0058] Existen diversas definiciones para el cálculo de la planicidad de un vector o la tonalidad de un espectro (que se relaciona inversamente con la planicidad de un espectro). La medida de planicidad espectral *SFM*

empleada aquí se calcula como la proporción del promedio geométrico y el promedio aritmético de los coeficientes espectrales *L* de la señal de sub-banda como se muestra en la Ecuación 7.

$$SFM(m,r) = \frac{e^{\left(\sum_{q=l_r}^{u_r} \log(|X(m,q)|)\right)/L}}{\frac{1}{L}\sum_{q=l_r}^{u_r} |X(m,q)|} \quad (7)$$

Asimetría espectral

[0059] La asimetría de una distribución mide su asimetría alrededor de su centroide y se define como el tercer momento central de una variable aleatoria dividido por el cubo de su desviación estándar.

Coeficientes de Predicción Lineal

[0060] Los Coeficientes de Predicción Lineal (LPC = Linear Prediction Coefficients) son los coeficientes de un filtro omnipolar, que pronostica el valor actual *x(k)* de una serie de tiempo a partir de valores anteriores, de tal manera que la media cuadrática del error $E = \sum_k (\hat{x}_k - x_k)^2$ se minimiza.

$$\hat{x}(k) = -\sum_{j=1}^{p} a_j x_{k-j} \quad (8)$$

[0061] Los LPC se calculan mediante el procedimiento de autocorrelación.

Coeficientes cepstral de frecuencia Mel

[0062] Los espectros de potencia son combados según la escala Mel mediante el uso de funciones de ponderación triangular con peso único para cada banda de frecuencia. MFCC se calculan al tomar el logaritmo y calcular la Transformada Coseno Discreto.

Coeficientes de predicción lineal perceptual de espectro relativos

[0063] Los coeficientes RASTA-PLP [H. Hermansky, N. Morgan, "RASTA Processing of Speech", IEEE Trans. On Speech and Audio Processing, vol. 2, no. 4, pp. 578-589, 1994] se calculan a partir de espectros de potencia en las siguientes etapas:

1. Compresión de magnitud de los coeficientes espectrales
2. Filtrado de paso de banda de la energía sub-banda con el tiempo
3. Expansión de magnitud que se relaciona con el procesamiento inverso de la etapa 2
4. Multiplicación con pesos que corresponden a una curva de volumen o nivel de ruido igual
5. Simulación de sensación de volumen o nivel de ruido al elevar los coeficientes a la potencia de 0,33
6. Cálculo de un modelo omnipolar de espectro resultante mediante el procedimiento de autocorrelación

Coeficientes de predicción lineal perceptual (PLP = Perceptual Linear Prediction)

[0064] Los valores PLP se calculan similares a los RASTA-PLP pero sin aplicar las etapas 1-3 [H. Hermansky, "Perceptual Linear Predictive Analysis for Speech", J. Ac. Soc. Am., vol. 87, no. 4, pp. 1738 - 1752, 1990].

Características delta

[0065] Características delta se han aplicado con éxito en reconocimiento de habla automático y clasificación de contenido de audio en el pasado. Existen diversas formas para su cálculo. Aquí, se calcula mediante convolución de la secuencia de tiempo de una característica con una pendiente lineal con una longitud de 9 muestras (la velocidad de muestreado de la serie de tiempo característica es igual a la velocidad de cuadro de STFT). Características delta-delta se obtienen al aplicar la operación delta a las características delta.

[0066] Como se ha indicado anteriormente, se prefiere tener una separación de banda de la banda de frecuencia de baja-resolución, que es similar a la situación perceptual del sistema de audición humana. Por lo tanto, se prefiere una separación de banda logarítmica o una separación de banda tipo Bark. Esto significa que las bandas que tienen una baja frecuencia central son más estrechas que las bandas que tienen una alta frecuencia central. En el cálculo de la medida de planicidad espectral, por ejemplo las operaciones de suma se extienden desde un valor q, que es normalmente el valor de frecuencia más bajo en una banda y se extienden al valor de cuenta u_r, que es el valor espectral más alto dentro de una banda predefinida. Para tener una mejor medida de planicidad espectral, se prefiere utilizar, en las bandas inferiores, al menos algunos o todos los valores espectrales de la banda de frecuencia adyacente inferior y/o superior. Esto significa que, por ejemplo la medida de planicidad espectral para la segunda banda, se calcula mediante el uso de los valores espectrales de la segunda banda y adicionalmente, mediante el uso de los valores espectrales de la primera banda y/o la tercera banda. En la realización preferida, no solo se emplean los valores espectrales de cualquiera de la primera o la segunda bandas, sino también se emplean los valores espectrales de la primera banda y la tercera banda. Esto significa que cuando se calcula SFM para la segunda banda, q en la Ecuación (7) se extiende desde l_r igual al primer valor espectral (más bajo) de la primera banda y u_r es igual al más alto valor espectral en la tercera banda. De esta manera, una característica de forma espectral, que se basa en un número superior de valores espectrales, puede calcularse hasta un cierto ancho de banda en el cual el número de valores espectrales dentro de la propia banda es suficiente de tal manera que l_r y u_r indiquen valores espectrales de la misma banda de frecuencia de baja resolución.

[0067] Con respecto a los coeficientes de predicción lineal, que se extraen por el extractor de características, se prefiere ya sea utilizar el LPC a_j de la Ecuación (8) o los valores residuales/de error que quedan después de la optimización o cualquier combinación de los coeficientes y los valores de error tal como una multiplicación o una adición con un factor de normalización de tal manera que los coeficientes así como los valores de error al cuadrado influencian la característica LPC extraída por el extractor de características.

[0068] Una ventana de la característica de forma espectral es que es una característica de baja-dimensión. Cuando, por ejemplo, se considera el ancho de banda de frecuencia que tiene 10 valores espectrales complejos o reales, el uso de todos estos 10 valores espectrales complejos o reales no sería útil y no sería un desperdicio de recursos computacionales. Por lo tanto, se extrae la característica de forma espectral, que tiene una dimensión, que es menor que la dimensión de los datos en bruto. Cuando por ejemplo, se considera la energía, entonces los datos en bruto tienen una dimensión de 10, ya que existen 10 valores espectrales al cuadrado. A fin de extraer la característica de forma espectral, que puede ser empleada eficientemente, se extrae una característica de forma espectral, que tiene una dimensión menor que la dimensión de los datos en bruto y que de preferencia, está en 1 ó 2. Una reducción de dimensión similar con respecto a los datos en bruto puede obtenerse cuando por ejemplo, se realiza un ajuste polinomio de bajo nivel a una envolvente espectral de una banda de frecuencia. Cuando por ejemplo, solo dos o tres parámetros se ajustan, entonces la característica de forma espectral incluye estos dos o tres parámetros de un polinomio o cualquier otro sistema de parametrización. En general, todos los parámetros, que indican la distribución de energía dentro de una banda de frecuencia y que tienen una baja dimensión menor al 5% o al menos menor que el 50% o solo menor que el 30% de la dimensión de datos en crudo o en bruto, son útiles.

[0069] Se ha encontrado que el uso de la característica de forma espectral solo ya resulta en un comportamiento ventajoso del aparato para procesar una señal de audio, pero se prefiere utilizar al menos una característica a modo de banda adicional. También se ha mostrado que la característica a modo de banda adicional útil para proporcionar resultados mejorados es la energía espectral por banda, que se calcula por cada cuadro de tiempo y banda de frecuencia y normaliza por la energía total del cuadro. Esta característica puede filtrarse de paso bajo o no. De forma adicional, se ha encontrado que la adición de la característica de flujo espectral mejora ventajosamente el desempeño del aparato de la invención, de tal manera que se obtiene un procedimiento eficiente que resulta en un buen desempeño cuando la característica de forma espectral por bandas se utiliza además de la característica de energía espectral por banda y la característica de flujo espectral por banda. Además de las características adicionales, esto mejora de nuevo el desempeño del aparato de la invención.

[0070] Como se discute con respecto a la característica de energía espectral, puede aplicarse un filtrado de paso bajo de esta característica con el tiempo o aplicar una normalización promedio de movimiento sobre el tiempo, pero no necesariamente debe de aplicarse. En el caso anterior, un promedio por ejemplo de las cinco características de forma espectral anteriores para la banda correspondiente, se calculan y el resultado de este cálculo se utiliza como la característica de forma espectral para la banda actual en el cuadro actual. Este promediado, sin embargo también puede aplicarse de forma bi-direccional, de tal manera que para la operación de promediado, no solo se utilizan características del pasado, sino también características del "futuro" para calcular la característica actual.

[0071] Las Figuras 7 y 8 se discutirán posteriormente a fin de proporcionar la implementación preferida del extractor de características 14 como se ilustra en las Figuras 1, 2 ó 4. En una primera etapa, una señal de audio se presenta en pequeñas ventanas a fin de proporcionar un bloque de valores de muestreado de audio como se indica en la etapa 70. De preferencia, se aplica una superposición. Esto significa que una y la misma muestra de audio ocurre en dos cuadros sucesivos debido al intervalo de superposición, en donde se prefiere una superposición del 50% con respecto a los valores de muestreado de audio. En la etapa 71, una conversión de tiempo/frecuencia de un bloque de valores de muestreado de audio presentados en pequeñas ventanas se realiza a fin de obtener una representación de frecuencia con una primera resolución, que es una alta resolución. Para este fin, se obtiene una transformada Fourier de corto tiempo (STFT = Short-Time Fourier Transform) con una FFT eficiente implementada. Cuando la etapa 71 se aplica varias veces con bloques temporalmente sucesivos de valores de muestreado de audio, un espectrograma se obtiene como se conoce en la técnica. En la etapa 72, la información espectral de alta resolución, es decir los valores espectrales de alta-resolución se agrupan en bandas de frecuencia de baja-resolución. Cuando por ejemplo, una FFT con 1024 ó 2048 valores de alimentación se aplica, 1024 ó 2048 valores espectrales existen, pero esta alta resolución ni se requiere ni se pretende. Por el contrario, la etapa de agrupamiento 72 resulta en una división del espectro de alta resolución en un pequeño número de bandas, tales como bandas que tienen un ancho de banda variante tal como por ejemplo conocido de las bandas Bark, o de una división de banda logarítmica. A continuación, subsecuente a la etapa de agrupamiento 72, un cálculo 73 de la característica de forma espectral y de preferencia otras características, se realiza para cada una de las bandas de baja resolución. Aunque no se indica en la Figura 7, características adicionales referentes a toda la banda de frecuencia pueden calcularse mediante el uso de los datos obtenidos en la etapa 70, ya que para estas características de ancho de banda íntegro, cualesquiera separaciones espectrales obtenidas por la etapa 71 o la etapa 72 no se requieren.

[0072] La etapa 73 resulta en características de forma espectral, que tienen m dimensiones, en donde m es menor que n y de preferencia es 1 ó 2 por banda de frecuencia. Esto significa que la información para una banda de frecuencia presente después de la etapa 72, se comprime en una información de baja dimensión presente después de la etapa 73 por la operación de extractor de características.

[0073] Como se indica en la Figura 7 cerca de la etapa 71 y la etapa 72, la etapa de conversión de tiempo/frecuencia y agrupamiento puede reemplazarse para diferentes operaciones. La salida de la etapa 70 puede filtrarse con un banco de filtro de baja-resolución que, por ejemplo se implementa, de tal manera que en la salida se obtienen 25 señales de sub-banda. El análisis de alta-resolución de cada sub-banda puede realizarse entonces para obtener los datos primarios para el cálculo de características de forma espectral. Esto puede realizarse, por ejemplo por un análisis FFT de una señal de sub-banda o por cualquier otro análisis de una señal de sub-banda, tal como por adicionales bancos de filtro en cascada.

[0074] La Figura 8 ilustra el procedimiento preferido para implementar el filtro controlable 12 de la Figura 1 o la característica de ponderación espectral ilustrada en la Figura 3 o indicada en 12 en la Figura 4. Subsecuente a la etapa de determinación de la información de control a modo de banda de baja resolución, tal como los valores SNR de sub-banda, que se envían de salida por el bloque de regresión de red neural 15 de la Figura 4, como se indica en la etapa 80, se realiza una interpolación lineal a la alta resolución en la etapa 81.

[0075] El propósito es obtener finalmente un factor de ponderación para cada valor espectral que se obtiene por la transformada Fourier de corto-tiempo realizada en la etapa 30 de la Figura 3, realizado en la etapa 71 o el procedimiento alterno indicado a la derecha de las etapas 71 y 72. Subsecuente a la etapa 81, se obtiene un valor SNR para cada valor espectral. Sin embargo, este valor SNR todavía está en el dominio logarítmico y en la etapa 82 proporciona una transformación del dominio logarítmico en un dominio lineal por cada valor espectral de alta-resolución.

[0076] En la etapa 83, los valores SNR lineales por cada valor espectral, es decir a la alta resolución son alisados con el tiempo y frecuencia, tal como al utilizar filtros de paso bajo IIR o de forma alterna, filtros de paso bajo FIR, por ejemplo puede aplicarse cualquier operación de promedio en movimiento. En la etapa 84, los pesos espectrales por cada uno de los valores de frecuencia de alta-resolución se calculan con base en los valores SNR lineales alisados. Este cálculo se basa en la función indicada en la Figura 5, aunque la función indicada en esta figura se da en términos logarítmicos, mientras que los pesos espectrales por cada valor de frecuencia de alta-resolución en la etapa 84 se calculan en el dominio lineal.

[0077] En la etapa 85, cada valor espectral se multiplica entonces por el peso espectral determinado para

obtener un conjunto de valores espectrales de alta-resolución, que se ha multiplicado por el conjunto de pesos espectrales. Este espectro procesado se convierte en frecuencia-tiempo en la etapa 86. Dependiendo del escenario de aplicación y dependiendo de la superposición empleado en la etapa 80, una operación de disminución cruzada puede realizarse entre dos bloques de valores de muestreado de audio con dominio de tiempo que se obtienen por dos etapas de conversión subsecuentes de frecuencia-tiempo, para atender artefactos de bloqueo.

[0078] Puede aplicarse una presentación en pequeñas ventanas adicional para reducir artefactos de convolución.

[0079] El resultado de la etapa 86 es un bloque de valores de muestreado de audio, que tiene un desempeño de habla mejorado, es decir el habla puede ser percibida mejor que si se compara con la señal de alimentación de audio correspondiente, en donde la mejora de habla no se ha realizado.

[0080] Dependiendo de ciertos requisitos de implementación de los procedimientos de la invención, los procedimientos de la invención pueden ser implementados en hardware o en software. La implementación puede ser realizada mediante el uso de un medio de almacenamiento digital, en particular, un disco, un DVD o un CD que tiene señales de control legibles electrónicamente almacenadas en el mismo, que cooperan con sistemas informáticos programables, de tal manera que se realicen los procedimientos de la invención. En general, la presente invención es por lo tanto un producto de programa informático con un código de programa almacenado en un soporte legible por máquina, el código de programa se opera para realizar los procedimientos de la invención cuando el producto de programa informático se ejecuta en un ordenador. En otras palabras, los procedimientos de la invención son por lo tanto un programa informático que tiene un código de programa para realizar al menos uno de los procedimientos de la invención cuando el programa informático se ejecuta en un ordenador.

[0081] Las realizaciones descritas son solamente ilustrativas para los principios de la presente invención. Se entiende que modificaciones y variaciones de los arreglos y detalles aquí descritos serán evidentes para otros expertos en la técnica. Es la intención, por lo tanto, estar limitados solo por el alcance de las reivindicaciones de patente pendientes y no por los detalles específicos presentados a modo de descripción y explicación de las presentes realizaciones.

REIVINDICACIONES

1. Aparato para el procesamiento de una señal de audio para obtener información de control por sub-banda para un filtro de mejora de habla, que comprende:

un extractor de características para la obtención de una secuencia de tiempo de representaciones espectrales de tiempo corto de la señal de audio y para la extracción de al menos una primera característica en cada banda de frecuencia de una pluralidad de bandas de frecuencia para una pluralidad de representaciones espectrales de tiempo corto, representando al menos una primera característica una forma espectral de una representación espectral de tiempo corto en una banda de frecuencia de la pluralidad de bandas de frecuencia,
en el que el extractor de características es operativo para extraer características secundarias adicionales que representan una característica de una representación espectral de corto tiempo por banda de frecuencia que es al menos una energía espectral, un flujo espectral entre marcos sucesivos y características delta o delta-delta y en el que el extractor de características es operativo para extraer adicionalmente características terceras para el ancho de banda completo que es al menos una de las características de LPC, incluyendo las características de LPC una señal de error de LPC, coeficientes de predicción lineal hasta un orden predefinido o una combinación de las señales de error de LPC y coeficientes de predicción lineal, coeficientes PLP, coeficientes RASTA-PLP, coeficientes cepstral de frecuencia Mel y características delta; y
un combinador de características para la combinación de al menos una primera característica, al menos una segunda característica y al menos una tercera característica mediante el uso de parámetros de combinación para obtener la información de control para el filtro de mejora de habla para una porción de tiempo de la señal de audio, en el que el combinador de características es operativo para combinar al menos una primera característica para cada banda de frecuencia que representa una forma espectral, al menos una segunda característica adicional por banda de frecuencia y al menos una tercera característica adicional para el ancho de banda completo mediante el uso de los parámetros de combinación.

2. Aparato según la reivindicación 1, en el que el extractor de características es operativo para aplicar una operación de conversión de frecuencia, en el que para una secuencia de instantes de tiempo, se obtiene una secuencia de representaciones espectrales, teniendo las representaciones espectrales bandas de frecuencia con anchos de banda no-uniformes, volviéndose un ancho de banda más grande con una frecuencia central incrementada de una banda de frecuencia.

3. Aparato según la reivindicación 1, en el que el extractor de características es operativo para calcular, como la primera característica, una medida de planicidad espectral por banda que representa una distribución de energía dentro de la banda, o como una segunda característica, una medida de energía normalizada por banda, estando basada la normalización en la energía total de un marco de señal, del cual se deriva la representación espectral, y
en el que el combinador de características es operativo para utilizar la medida de planicidad espectral para una banda o la energía normalizada por banda.

4. Aparato según una de las reivindicaciones anteriores, en el que el extractor de características es operativo para extraer adicionalmente como la segunda característica, para cada banda, una medida de flujo espectral que representa una similaridad o diferencia entre representaciones espectrales sucesivas en tiempo o una medida de asimetría espectral, representando la medida de asimetría espectral una asimetría alrededor de un centroide.

5. Aparato según la reivindicación 1, en el que el extractor de características es operativo para calcular las características de coeficiente de predicción lineal para un bloque de muestras de audio de dominio de tiempo, incluyendo el bloque muestras de audio utilizadas para extraer la característica como mínimo que representa la forma espectral por cada banda de frecuencia.

6. Aparato según la reivindicación 1, en el que el extractor de características es operativo para calcular la forma del espectro en una banda de frecuencia mediante el uso de información espectral de una o dos bandas de frecuencia inmediatamente adyacentes y la información espectral de la banda de frecuencia solo.

7. Aparato según la reivindicación 1, en el que el extractor de características es operativo para extraer información de características en bruto por cada característica por bloque de muestras de audio y combinar la secuencia de información de características en bruto en una banda de frecuencia para obtener al menos una primera característica por la banda de frecuencia.

8. Aparato según la reivindicación 1, en el que el extractor de características es operativo para calcular, por cada banda de frecuencia, un número de valores espectrales y para combinar el número de valores espectrales, para obtener al menos una primera característica que representa la forma espectral de tal manera que al menos una primera característica tiene una dimensión, que es más pequeña que el número de valores espectrales en la banda de frecuencia.

9. Procedimiento de procesamiento de una señal de audio para obtener información de control por sub-banda para un filtro de mejora de habla, que comprende:

la obtención de una secuencia de tiempo de representaciones espectrales de corto tiempo de la señal de audio,
la extracción de al menos una primera característica en cada banda de frecuencia de una pluralidad de bandas de frecuencia para una pluralidad de representaciones espectrales de corto tiempo, representando al menos una primera característica una forma espectral de una representación espectral de corto tiempo en una banda de frecuencia de la pluralidad de bandas de frecuencia,
en el que segundas características adicionales que representan una característica de una representación espectral de corto tiempo por banda de frecuencia son al menos una de una energía espectral, un flujo espectral entre marcos sucesivos y se extraen características delta o delta-delta y en el que terceras características adicionales para el ancho de banda completo son al menos una de las características de LPC, incluyendo las características de LPC una señal de error de LPC, coeficientes de predicción lineal hasta un orden predefinido o una combinación de las señales de error de LPC y coeficientes de predicción lineal, coeficientes PLP, coeficientes RASTA-RLP, coeficientes cepstral de frecuencia Mel y características delta son extraídos; y
la combinación de al menos una primera característica, al menos una segunda característica y al menos una tercera característica mediante el uso parámetros de combinación para obtener la información de control para el filtro de mejora de habla para una porción de tiempo de la señal de audio, en el que al menos una primera característica para cada banda de frecuencia que representa una forma espectral, al menos una segunda característica adicional por banda de frecuencia y al menos una tercera característica adicional para el ancho de banda completo se combinan para cada banda de frecuencia mediante el uso de los parámetros de combinación.

10. Aparato para mejora de habla en una señal de audio, que comprende:

un aparato para el procesamiento de la señal de audio para obtener información de control para un filtro de mejora de habla para una pluralidad de bandas que representan una porción de tiempo de la señal de audio según la reivindicación 1; y
un filtro controlable, siendo controlable el filtro de tal manera que una banda de la señal de audio se atenúa de forma variable con respecto a una banda diferente con base en la información de control.

11. Aparato según la reivindicación 10, en el que el aparato para procesamiento incluye el convertidor de frecuencia-tiempo que proporciona información espectral que tiene una resolución superior que una resolución espectral, para lo cual se proporciona información de control; y
en el que el aparato comprende adicionalmente un post-procesador de información de control para interpolar la información de control a la alta resolución y para alisar la información de control interpolada, para obtener una información de control post-procesada con base en que parámetros de filtro controlables del filtro controlable se ajustan.

12. Procedimiento para mejora de habla en una señal de audio, que comprende:

un procedimiento de procesamiento de la señal de audio para obtener información de control para un filtro de mejora de habla para una pluralidad de bandas que representan una porción de tiempo de la señal de audio según la reivindicación 9; y
el control de un filtro de tal manera que una banda de la señal de audio se atenúa de forma variable con respecto a una banda diferente con base en la información de control.

13. Aparato para entrenar un combinador de características, para determinar parámetros de combinación del combinador de características, que comprende:

un extractor de características para obtener una secuencia de tiempo de representaciones espectrales de corto tiempo de una señal de audio de entrenamiento, para lo cual se conoce una información de control para el filtro de mejora de habla por banda de frecuencia, y para extraer al menos una característica en cada banda de frecuencia de la pluralidad de bandas de frecuencia para una pluralidad de representaciones espectrales de corto tiempo,

representando al menos una primera característica una forma espectral de una representación espectral de corto tiempo en una banda de frecuencia de la pluralidad de bandas de frecuencia;
en el que el extractor de características es operativo para extraer características secundarias adicionales que representan una característica de una representación espectral de corto tiempo por banda de frecuencia que es al menos una energía espectral, un flujo espectral entre marcos sucesivos y características delta o delta-delta y en el que el extractor de características es operativo para extraer adicionalmente características terceras para el ancho de banda completo que es al menos una de las características de LPC, incluyendo las características de LPC una señal de error de LPC, coeficientes de predicción lineal hasta un orden predefinido o una combinación de las señales de error de LPC y coeficientes de predicción lineal, coeficientes PLP, coeficientes RASTA-PLP, coeficientes cepstral de frecuencia Mel y características delta; y un controlador de optimización para alimentar el combinador de características con al menos una primera característica, al menos una segunda característica y al menos una tercera característica por cada banda de frecuencia, para calcular la información de control mediante el uso de parámetros de combinación intermedios, para variar los parámetros de combinación intermedios, para comparar la información de control variada con la información de control conocida y para actualizar los parámetros de combinación intermedios, cuando los parámetros de combinación intermedios variados resultan en información de control que se corresponde mejor con la información de control conocida, en el que el combinador de características es operativo para combinar al menos una característica para cada banda de frecuencia que representa una forma espectral, al menos una segunda característica adicional por banda de frecuencia y al menos una tercera característica adicional para el ancho de banda completo mediante el uso de los parámetros de combinación.

14. Procedimiento para entrenar un combinador de características, para la determinación de parámetros de combinación del combinador de características, que comprende:

la obtención de una secuencia de tiempo de representaciones espectrales de corto tiempo de una señal de audio de entrenamiento, para lo cual se conoce una información de control para un filtro de mejora de habla por banda de frecuencia;
la extracción de al menos una primera característica en cada banda de frecuencia de la pluralidad de bandas de frecuencia para una pluralidad de representaciones espectrales de corto tiempo, representando al menos una primera característica una forma espectral de una representación espectral de corto tiempo en una banda de frecuencia de la pluralidad de bandas de frecuencia;
en el que las segundas características adicionales que representan una característica de una representación espectral de corto tiempo por banda de frecuencia que es al menos una de una energía espectral, un flujo espectral entre marcos sucesivos, y características delta o delta-delta son extraídas, y
en el que terceras características adicionales para el ancho de banda completo son al menos una de las características de LPC, incluyendo las características de LPC una señal de error de LPC, coeficientes de predicción lineal hasta un orden predefinido o una combinación de las señales de error de LPC y coeficientes de predicción lineal, coeficientes PLP, coeficientes RASTA-RLP, coeficientes cepstral de frecuencia Mel y características delta son extraídos;
la alimentación del combinador de características con al menos una primera, al menos una segunda y al menos una tercera característica por cada banda de frecuencia;
el cálculo de la información de control mediante el uso de parámetros de combinación intermedios;
la variación de los parámetros de combinación intermedios;
la comparación de la información de control variada con la información de control conocida;
la actualización de los parámetros de combinación intermedios, cuando los parámetros de combinación intermedios variados resultan en información de control que se corresponde mejor con la información de control conocida,
en el que el combinador de características es operativo para combinar al menos una primera característica para cada banda de frecuencia que representa una forma espectral, al menos una segunda característica adicional por banda de frecuencia y al menos una tercera característica adicional para el ancho de banda completo mediante el uso de parámetros de combinación.

15. Programa informático para realizar, cuando se ejecuta en un ordenador, un procedimiento según la reivindicación 9, 12 ó 14.

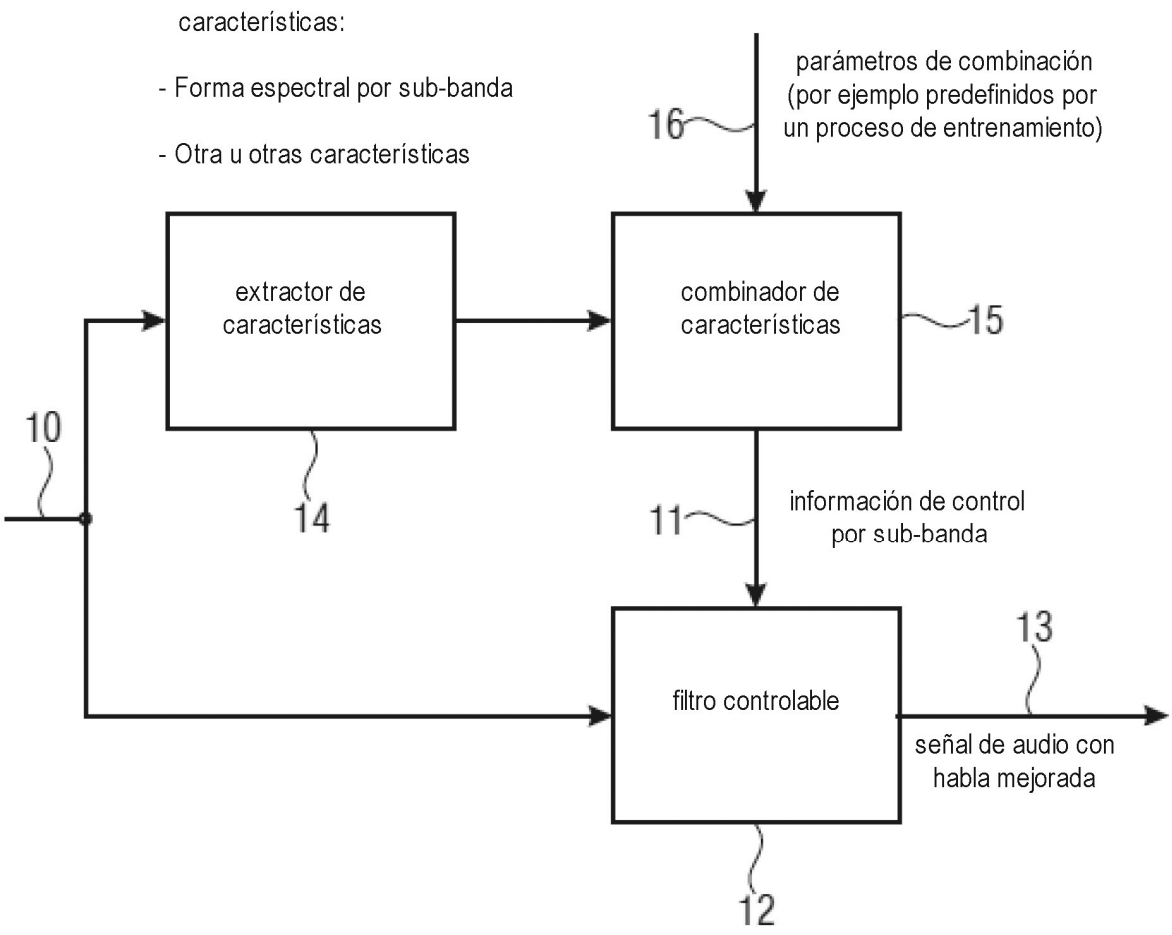

características:
- Forma espectral por sub-banda
- Otra u otras características
parámetros de combinación (por ejemplo predefinidos por un proceso de entrenamiento)
16
extractor de características
combinador de características
15
10
14
11
información de control por sub-banda
filtro controlable
13
señal de audio con habla mejorada
12

FIGURA 1

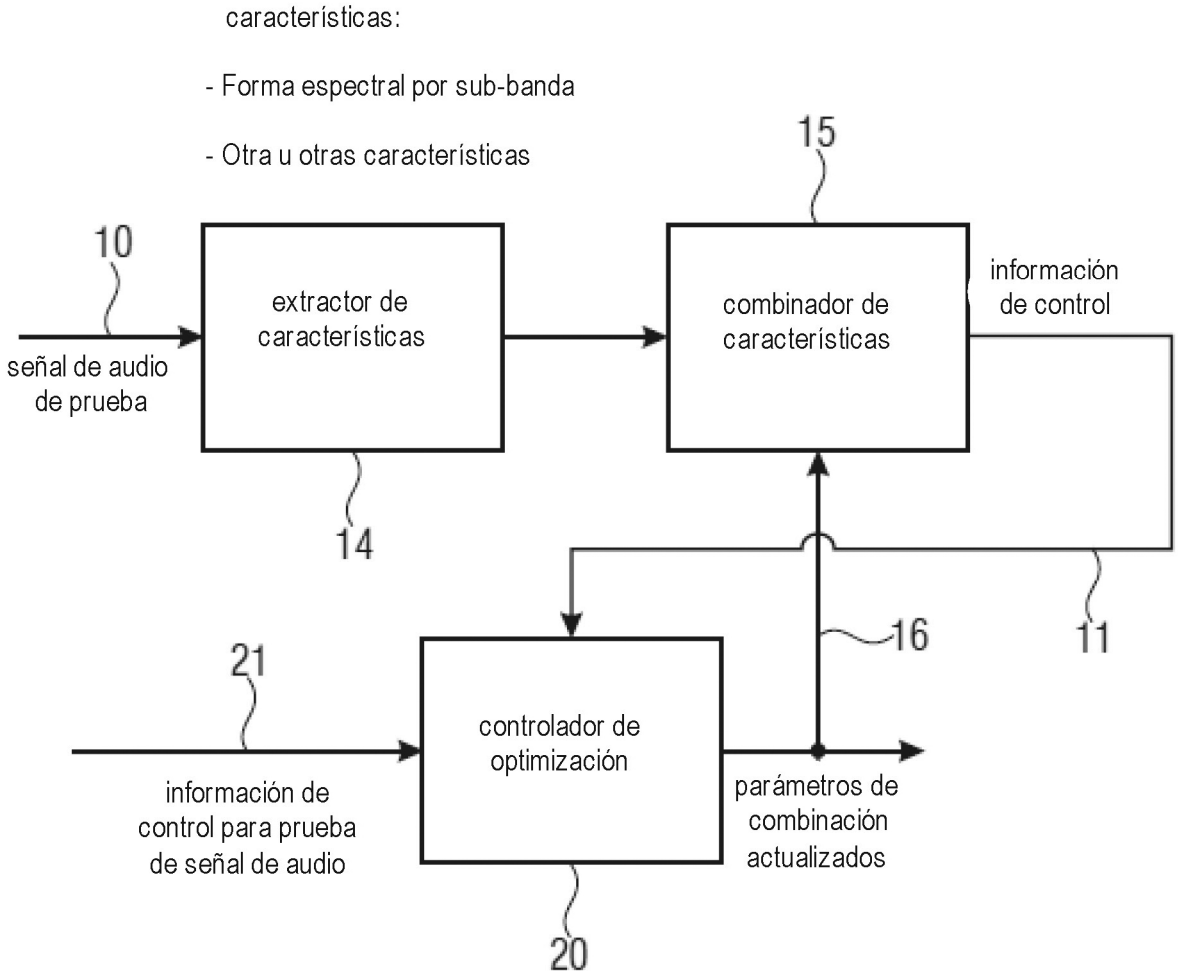
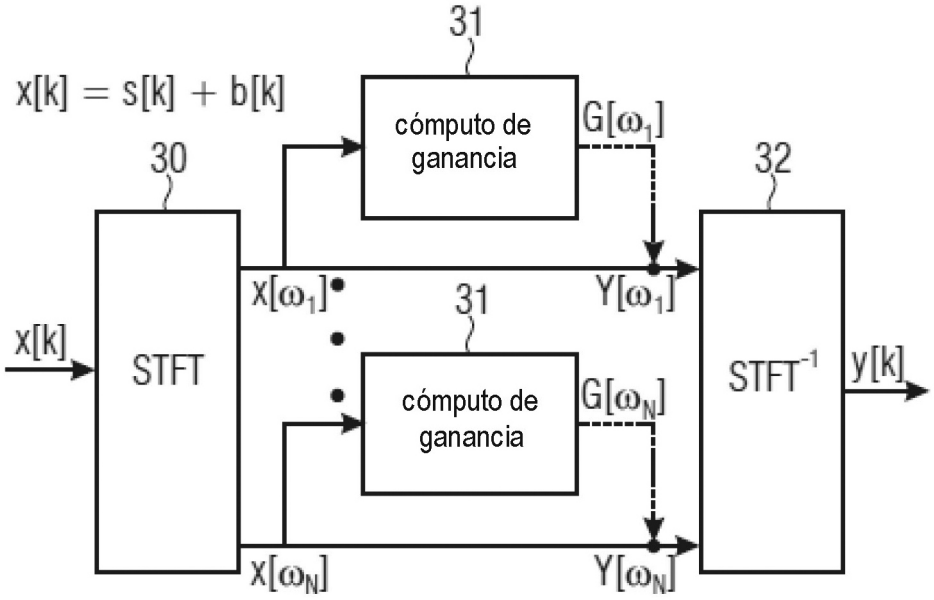

características:
- Forma espectral por sub-banda
- Otra u otras características
15
10
extractor de características
combinador de características
información de control
señal de audio de prueba
14
16
11
21
controlador de optimización
información de control para prueba de señal de audio
parámetros de combinación actualizados
20

FIGURA 2



$$Y(\omega) = G(\omega)X(\omega)$$

$$G(\omega) = \sqrt{1 - \frac{|\hat{B}(\omega)|^2}{|X(\omega)|^2}}$$

$$G(\omega) = \sqrt{\frac{\hat{R}(\omega)}{\hat{R}(\omega) + 1}}$$

información de control puede ser:

-SNR ($\hat{R}(\omega)$) por banda

-energía de habla $\hat{S}(\omega)$ por banda

-energía de "ruido" de fondo $\hat{B}(\omega)$ por banda

-parámetros de filtro para que el filtro controlable obtenga filtrado deseado

FIGURA 3

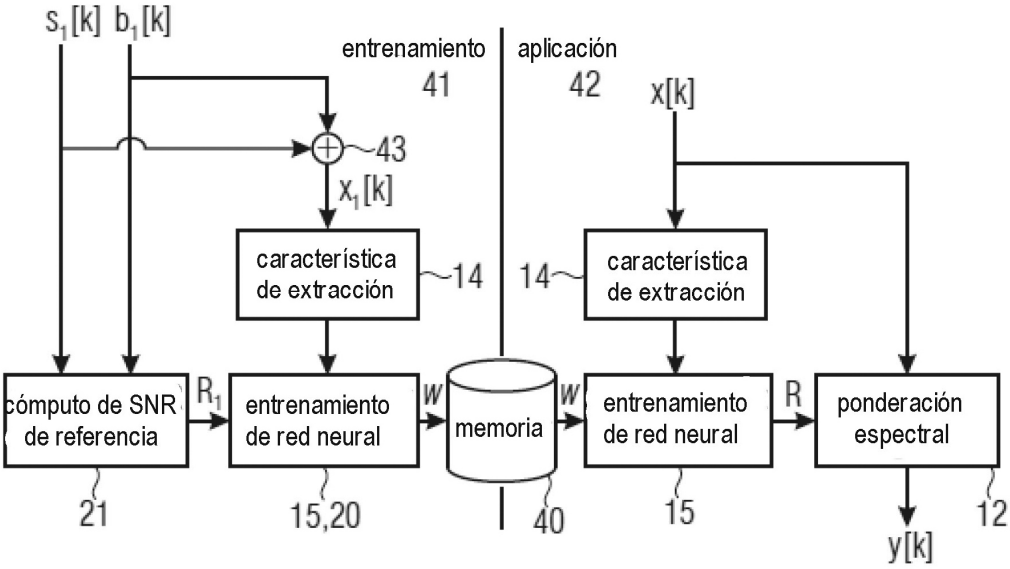
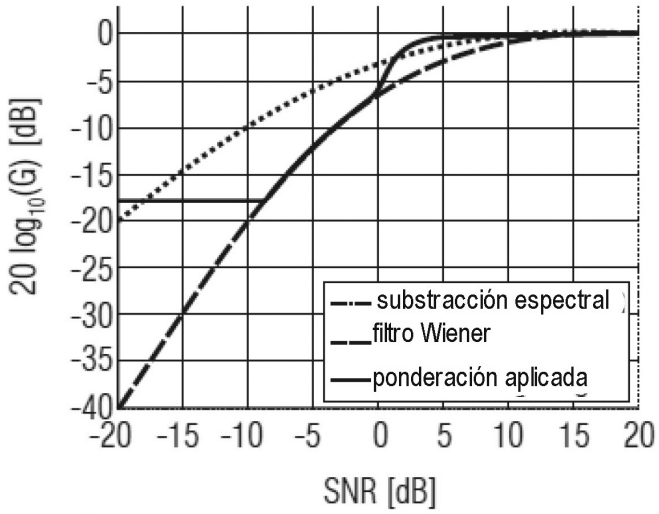


FIGURA 4



$$G(\omega) = \begin{cases} \dfrac{\hat{R}(\omega)^{\alpha}}{\hat{R}(\omega)^{\alpha}+1} & \hat{R}(\omega) \le 1 \\ \dfrac{\hat{R}(\omega)^{\beta}}{\hat{R}(\omega)^{\beta}+1} & \hat{R}(\omega) > 1 \end{cases}$$

e.g. $\alpha = 3{,}5$
$\beta = 1$
límite a -18dB

FIGURA 5

- características por banda de frecuencia (banda de ponderación, por ejemplo banda Bark)
  - energía espectral en banda
  - forma espectral en banda
    (distribución de energía dentro de la banda)
    - medida de planicidad espectral $\left(\frac{\text{promedio geométrico}}{\text{promedio aritmético}}\right)$
    - asimetría espectral
  - flujo espectral entre cuadros sucesivos
  - características delta o delta-delta

- características para todo el ancho de banda
  - coeficiente LPC y/o señal de error LPC
  - coeficientes cepstral de frecuencia Mel
  - coeficientes de predicción lineal perceptual de espectro relativo (RAST A-PLP)
  - características delta o delta-delta

FIGURA 6

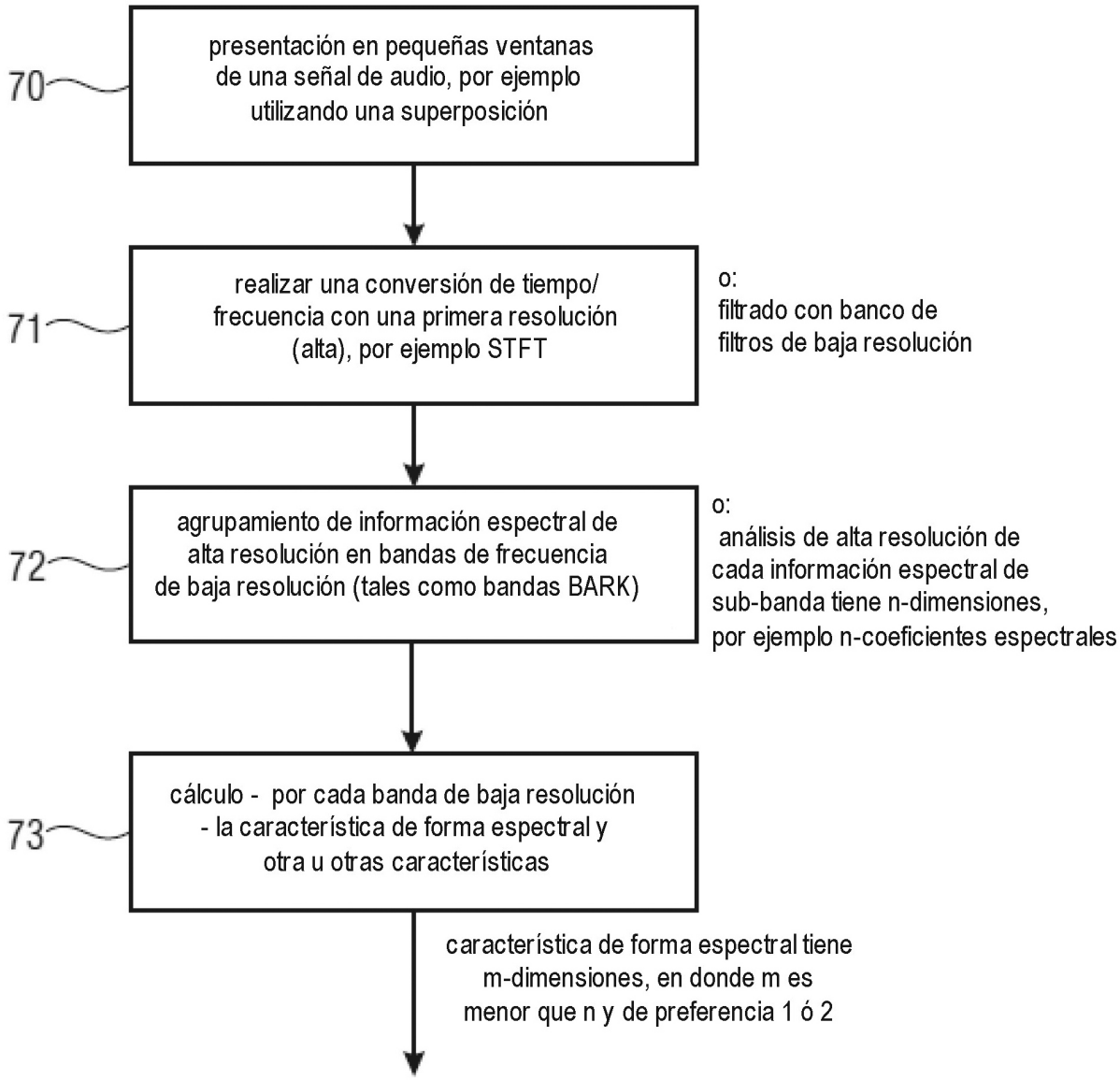

70
presentación en pequeñas ventanas de una señal de audio, por ejemplo utilizando una superposición
71
realizar una conversión de tiempo/ frecuencia con una primera resolución (alta), por ejemplo STFT
o: filtrado con banco de filtros de baja resolución
72
agrupamiento de información espectral de alta resolución en bandas de frecuencia de baja resolución (tales como bandas BARK)
o: análisis de alta resolución de cada información espectral de sub-banda tiene n-dimensiones, por ejemplo n-coeficientes espectrales
73
cálculo - por cada banda de baja resolución - la característica de forma espectral y otra u otras características
característica de forma espectral tiene m-dimensiones, en donde m es menor que n y de preferencia 1 ó 2

FIGURA 7

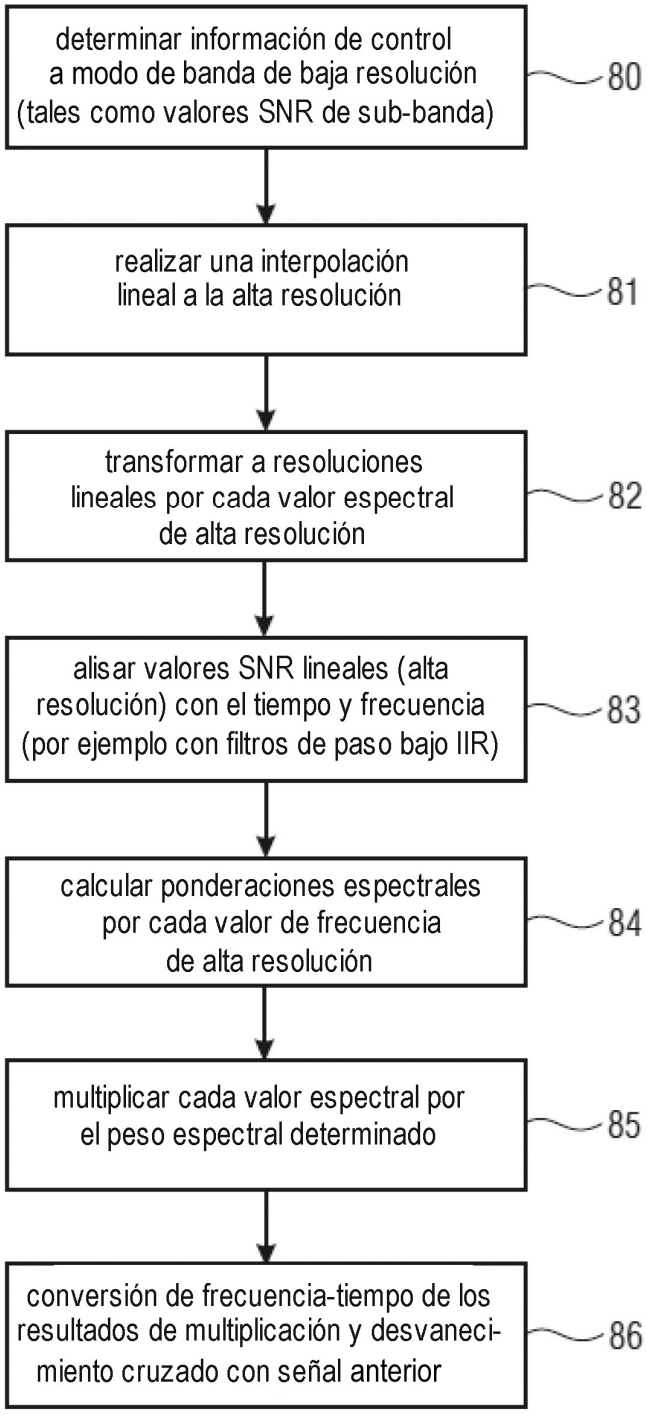

determinar información de control a modo de banda de baja resolución (tales como valores SNR de sub-banda)
80
realizar una interpolación lineal a la alta resolución
81
transformar a resoluciones lineales por cada valor espectral de alta resolución
82
alisar valores SNR lineales (alta resolución) con el tiempo y frecuencia (por ejemplo con filtros de paso bajo IIR)
83
calcular ponderaciones espectrales por cada valor de frecuencia de alta resolución
84
multiplicar cada valor espectral por el peso espectral determinado
85
conversión de frecuencia-tiempo de los resultados de multiplicación y desvaneci-miento cruzado con señal anterior
86

FIGURA 8

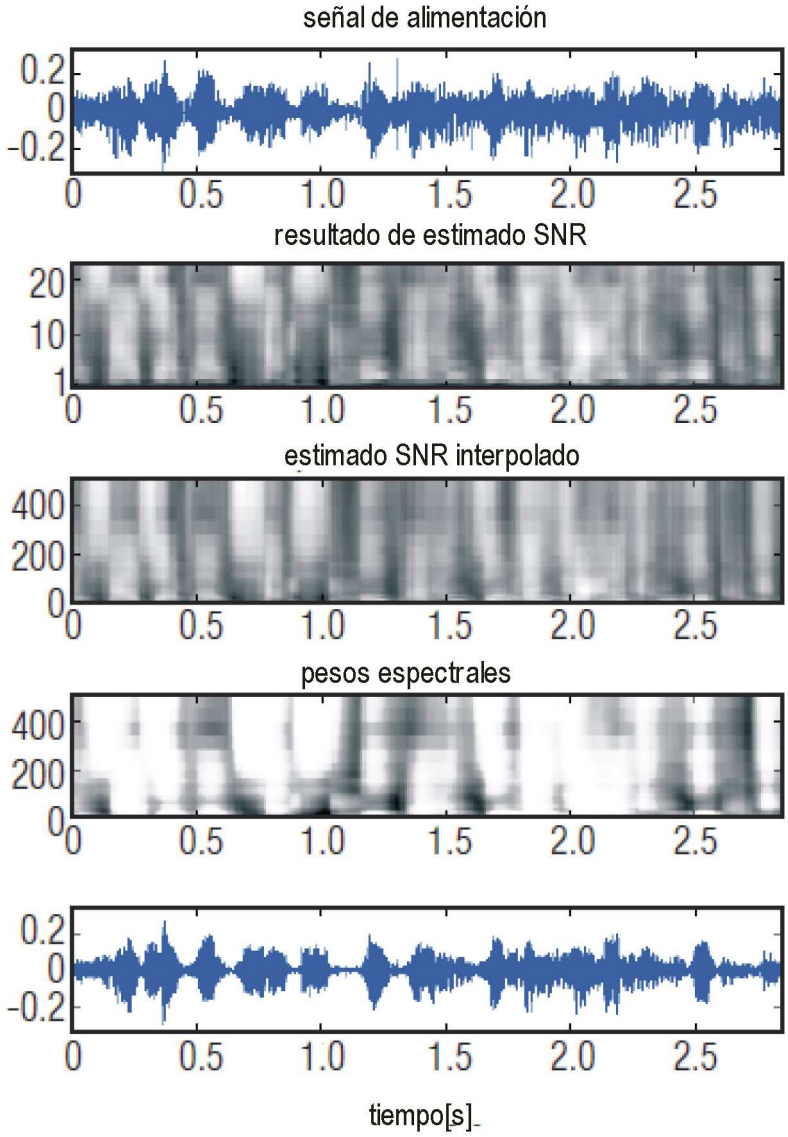

señal de alimentación
0.2
0
-0.2
0
0.5
1.0
1.5
2.0
2.5
resultado de estimado SNR
20
10
1
estimado SNR interpolado
400
200
0
pesos espectrales
tiempo[s]

FIGURA 9

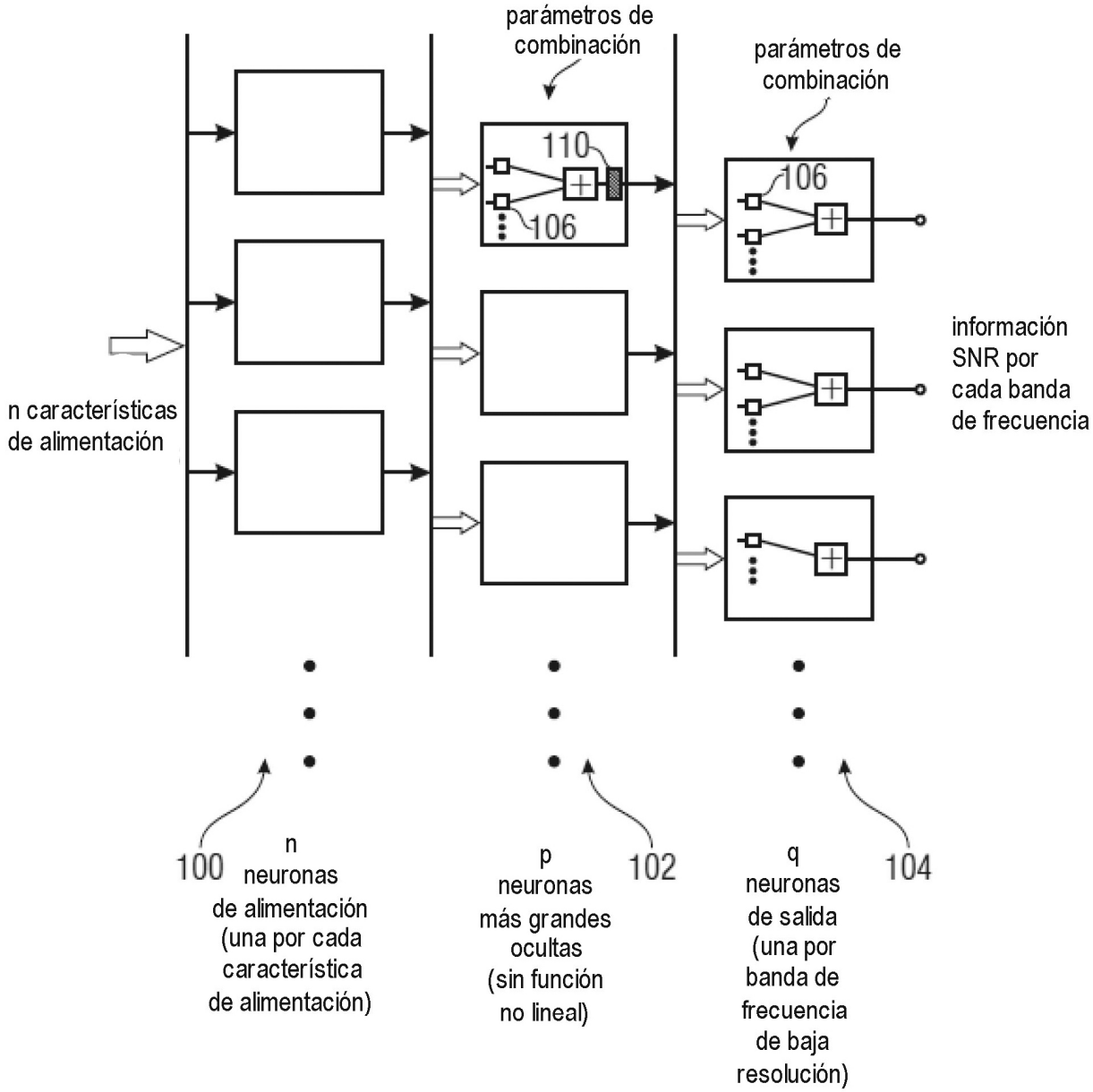

parámetros de combinación
parámetros de combinación
110
106
106
información SNR por cada banda de frecuencia
n características de alimentación
100
n neuronas de alimentación (una por cada característica de alimentación)
p neuronas más grandes ocultas (sin función no lineal)
102
q neuronas de salida (una por banda de frecuencia de baja resolución)
104

FIGURA 10