(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 364 888**

(51) Int. Cl.:
G10L 21/02 (2006.01)
H04S 5/02 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **08802737 .0**
(96) Fecha de presentación : **01.10.2008**
(97) Número de publicación de la solicitud: **2206113**
(97) Fecha de publicación de la solicitud: **14.07.2010**

(54) Título: **Dispositivo y procedimiento para generar una señal multicanal con un procesamiento de señal de voz.**

(30) Prioridad: **12.10.2007 DE 10 2007 048 973**

(45) Fecha de publicación de la mención BOPI:
16.09.2011

(45) Fecha de la publicación del folleto de la patente:
16.09.2011

(73) Titular/es: **Fraunhofer-Gesellschaft zur Förderung der Angewandten Forschung e.V.**
Hansastrasse 27C
80686 München, DE

(72) Inventor/es: **Uhle, Christian;**
Hellmuth, Oliver;
Herre, Jürgen;
Popp, Harald y
Kastner, Thorsten

(74) Agente: **Arizti Acha, Mónica**

DESCRIPCIÓN

Dispositivo y procedimiento para generar una señal multicanal con un procesamiento de señal de voz

La presente invención se refiere al campo del procesamiento de señales de audio y en particular a la generación de varios canales de salida a partir de menos canales de entrada, como por ejemplo un canal (mono) o dos canales (estéreo) de entrada.

Un dispositivo correspondiente se desprende por ejemplo del documento EP 1 021 063.

El material de audio multicanal se está volviendo cada vez más popular. Esto ha llevado a que, entretanto, también muchos usuarios finales tengan sistemas de reproducción multicanal. Esto se debe principalmente a que los DVD son cada vez más populares, y que por tanto también muchos usuarios de DVD entretanto tengan equipos multicanal 5.1. Los sistemas de reproducción de este tipo están compuestos en general por tres altavoces L (izquierdo), C (central) y R (derecho), que normalmente están dispuestos delante del usuario, y dos altavoces Ls y Rs, que están dispuestos detrás del usuario, y normalmente aún un canal LFE, que también se denomina canal de efecto de baja frecuencia o *subwoofer*. Un escenario de canales de este tipo se indica en la figura 5b y en la figura 5c. Mientras que el posicionamiento de los altavoces L, C, R, Ls, Rs respecto al usuario debería realizarse como se indica en las figuras 10 y 11, para que el usuario reciba una impresión auditiva lo mejor posible, el posicionamiento del canal LFE (no se muestra en las figuras 5b y 5c) no es tan decisivo porque a frecuencias tan bajas el oído no puede efectuar ninguna localización y así el canal LFE puede disponerse en cualquier lugar en el que no moleste por su tamaño considerable.

Un sistema multicanal de este tipo ofrece varias ventajas respecto a una reproducción estéreo típica, que es una reproducción de dos canales, tal como se muestra por ejemplo en la figura 5a. También fuera de la posición de escucha central óptima se obtiene una estabilidad mejorada de la impresión auditiva frontal, que también se denomina “imagen frontal” (*front image*), y concretamente gracias al canal central. Así se obtiene un “punto óptimo” (*sweet-spot*) más grande, indicando el “punto óptimo” la posición de escucha óptima.

Además el oyente tiene una mejor sensación de “inmersión” en la escena de audio gracias a los dos altavoces traseros Ls y Rs.

Aún así existe un gran número de materiales de audio en propiedad del usuario o disponibles en general que sólo existe como material estéreo, que por tanto sólo tiene dos canales, concretamente el canal izquierdo y el canal derecho. Los soportes de sonido típicos para tales piezas estéreo son los discos compactos.

Para reproducir un material estéreo de este tipo con un equipo de audio multicanal 5.1 existen dos opciones recomendadas por la ITU.

La primera opción consiste en reproducir el canal izquierdo y el derecho a través del altavoz izquierdo y el derecho del sistema de reproducción multicanal. Sin embargo, en esta solución es desventajoso que no se aproveche la pluralidad de altavoces ya existentes, es decir que no se aproveche la existencia del altavoz central y de los dos altavoces traseros.

Otra opción consiste en transformar los dos canales en una señal multicanal. Esto puede ocurrir durante la reproducción o mediante un procesamiento previo especial que aprovecha ventajosamente los seis altavoces del sistema de reproducción 5.1 existente por ejemplo y así lleva a una impresión auditiva mejorada cuando la mezcla ascendente o “*upmix*” se realiza sin errores de dos canales a 5 ó 6 canales.

La segunda opción, es decir, el uso de todos los altavoces del sistema multicanal sólo es ventajosa respecto a la primera solución cuando por tanto no se producen errores de mezcla ascendente. Tales errores de mezcla ascendente pueden ser en particular molestos cuando las señales para los altavoces traseros, que también se conocen como señales de ambiente o señales de entorno, no se generan sin errores.

Una posibilidad para realizar este denominado proceso de mezcla ascendente se conoce por la expresión “concepto directo/ambiente”. Las fuentes de sonido directo se reproducen mediante los tres canales frontales de manera que el usuario las percibe en la misma posición que en la versión original de dos canales. La versión original de dos canales se representa esquemáticamente en la figura 5a y concretamente en el ejemplo de diferentes instrumentos de percusión.

La figura 5b muestra una versión mezclada de manera ascendente del concepto, en la que todas las fuentes de sonido originales, esto es, los instrumentos de percusión se reproducen de nuevo por los tres altavoces L, C y R frontales, emitiéndose adicionalmente por los dos altavoces traseros señales de entorno especiales. La expresión “fuente de sonido directo” se utiliza así para describir un sonido que sólo procede directamente de una fuente de sonido discreta como por ejemplo un instrumento de percusión u otro instrumento o en general un objeto de audio especial, tal como se representa esquemáticamente por ejemplo en la figura 5a mediante un instrumento de percusión. Sonidos adicionales de cualquier tipo, como por ejemplo los debidos a reflexiones en la pared, etc. no están presentes en una fuente de sonido directo de este tipo. En este escenario las señales de sonido, emitidas por los dos altavoces traseros Ls, Rs en la figura 5b, sólo están compuestas por señales de entorno, que están presentes o no en la grabación original. Tales señales de entorno o señales “ambiente” no pertenecen a una única fuente de sonido, sino que contribuyen a la

reproducción de la acústica arquitectónica de una grabación y llevan por tanto a la denominada sensación de "inmersión" del oyente.

Un concepto alternativo adicional, denominado concepto "en la banda" se representa esquemáticamente en la figura 5c. Cualquier tipo de sonido, es decir, fuentes de sonido directo y sonidos de tipo de entorno se posiciona alrededor del oyente. La posición de un sonido es independiente de su característica (fuentes de sonido directo o sonidos de tipo de entorno) y depende sólo del diseño específico del algoritmo, tal como se representa por ejemplo en la figura 5c. Así en la figura 5c mediante el algoritmo de mezcla ascendente se determina que los dos instrumentos 1100 y 1102 se posicionan lateralmente respecto al oyente, mientras que los dos instrumentos 1104 y 1106 se posicionan delante del usuario. Esto lleva a que ahora los dos altavoces traseros Ls, Rs también contengan partes de los dos instrumentos 1100 y 1102 y ya no sólo sonidos de tipo de entorno, como aún era el caso de la figura 5b en el que los mismos instrumentos están posicionados todos delante del usuario.

La publicación "C. Avendano y J.M. Jot: "Ambience Extraction and Synthesis from Stereo Signals for Multichannel Audio Upmix", IEEE International Conference on Acoustics, Speech and signal Processing, ICASSP 02, Orlando, Fl, mayo de 2002" da a conocer una técnica en el dominio de frecuencia, para identificar y extraer información de entorno en señales de audio estéreo. Este concepto se basa en el cálculo de una coherencia entre canales y una función de correlación no lineal, que permitirá determinar regiones en tiempo-frecuencia en la señal estéreo que principalmente se componen de componentes de entorno. A continuación se sintetizan y utilizan señales de entorno para alimentar los canales traseros o canales "envolventes" Ls, Rs (figuras 10 y 11) de un sistema de reproducción multicanal.

En la publicación "R. Irwan y Ronald M. Aarts: "A method to convert stereo to multi-channel sound", The proceedings of the AES 19th International Conference, Schloss Elmau, Alemania, 21-24 de junio, páginas 139-143, 2001" se presenta un procedimiento para transformar una señal estéreo en una señal multicanal. La señal para los canales envolventes se calcula utilizando una técnica de correlación cruzada. Se utiliza un análisis de componente principal (PCA; PCA = *Principle Component Analysis*) para calcular un vector que indica una dirección de la señal dominante. Este vector se correlaciona entonces a partir de una representación de dos canales para dar una representación de tres canales con el fin de generar los tres canales frontales.

Todas las técnicas conocidas intentan extraer de diferentes maneras las señales de ambiente o señales de entorno de la señal estéreo original o incluso sintetizarlas a partir de ruidos o información adicional, pudiendo utilizarse para la síntesis de las señales de ambiente también información que no se encuentra en la señal estéreo. Sin embargo, en última instancia, siempre se trata de extraer información de la señal estéreo o alimentar información a un escenario de reproducción que no existe de manera explícita, ya que normalmente sólo está disponible una señal estéreo de dos canales y dado el caso alguna información adicional o metainformación.

A continuación se hará referencia a otros procedimientos de *upmix* o mezcla ascendente que funcionan sin parámetros de control. Tales procedimientos de mezcla ascendente se denominan también procedimientos de mezcla ascendente ciegos o procedimientos "*blind-upmixing*".

La mayor parte de tales técnicas para generar a partir de un monocanal una denominada señal pseudoestereofónica (esto es una mezcla ascendente de 1 a 2) no es adaptativa con respecto a la señal. Esto significa que siempre procesan igual una señal mono independientemente de qué contenido haya en la señal mono. Tales sistemas funcionan a menudo con retardos de tiempo y/o estructuras de filtro sencillas para descorrelacionar las señales generadas, por ejemplo mediante el procesamiento de la señal de entrada de un canal mediante un par de denominados filtros de peine complementarios, tal como se describe en M. Schroeder, "An artificial stereophonic effect obtained from using a single signal", JAES, 1957. Una visión general adicional de tales sistemas se encuentra en C. Faller, "Pseudo stereophony revisited", Proceedings of the AES 118nd Convention, 2005.

Además existe también la técnica de la extracción de la señal de entorno (*ambience extraction*) utilizando una factorización no negativa de matrices, en particular en el contexto de una mezcla ascendente de 1 a N, siendo N mayor que dos. En este caso se calcula una distribución de tiempo-frecuencia (TFD; TFD = *time-frequency distribution*) de la señal de entrada, por ejemplo por medio de una transformada de Fourier a corto plazo. Un valor de estimación de la TFD de las componentes de señal directa se deriva mediante un procedimiento de optimización numérico, que se denomina factorización no negativa de matrices. Un valor de estimación para la TFD de la señal de entorno se determina mediante el cálculo de la diferencia de la TFD de la señal de entrada y del valor de estimación de la TFD para la señal directa. La resíntesis o síntesis de la señal de tiempo de la señal de entorno se realiza utilizando el espectrograma de fase de la señal de entrada. Un procesamiento trasero adicional se realiza de manera opcional para mejorar la experiencia auditiva de la señal multicanal generada. Este procedimiento se describe detalladamente en C. Uhle, A. Walther, O. Hellmuth y J. Herre en "Ambience separation from mono recordings using non-negative matrix factorization", Proceedings of the AES 30th Conference 2007.

En la mezcla ascendente de grabaciones estéreo existen diferentes técnicas. Una técnica consiste en el uso de decodificadores de matriz. Los decodificadores de matriz se conocen con el nombre Dolby Pro Logic II, DTS Neo: 6 o HarmanKardon/Lexicon Logic 7 y están contenidos prácticamente en cualquier receptor de audio/vídeo que se venda hoy en día. Como subproducto de su funcionalidad prevista estos procedimientos también pueden realizar una mezcla ascendente ciega. Estos decodificadores usan diferencias entre canales y mecanismos de control adaptativos con

respecto a la señal para generar señales de salida multicanal.

Como ya se ha expuesto, también se utilizan técnicas en el dominio de frecuencia descritas por Avendano y Jot, para identificar y extraer la información de entorno (información de ambiente) en señales de audio estéreo. Este procedimiento se basa en el cálculo de un índice de coherencia entre canales y una función de correlación no lineal, por lo que es posible determinar las regiones de tiempo-frecuencia que se componen principalmente de componentes de señal de entorno. A continuación se sintetizan y utilizan las señales de entorno para alimentar los canales envolventes del sistema de reproducción multicanal.

Un elemento constitutivo del proceso de mezcla ascendente directo/de entorno consiste en la extracción de una señal de entorno que se alimenta a los dos canales traseros Ls, Rs. Existen determinados requisitos con respecto a una señal para su utilización como señal de tipo de entorno en el contexto de un proceso de mezcla ascendente directo/de entorno. Una condición previa consiste en que no deben ser audibles partes relevantes de las fuentes de sonido directo para poder situar las fuentes de sonido directo de manera segura delante del oyente. Esto es en particular importante cuando la señal de audio contiene voz o uno o más hablantes distinguibles. En cambio, las señales de voz generadas por una multitud no tienen por qué molestar obligatoriamente al oyente si no están situadas delante del oyente.

Si se reprodujera una cantidad especial de componentes de voz a través de los canales traseros esto llevaría a que la posición del o de los pocos hablantes se situara de delante hacia atrás o algo más lejos del usuario o incluso detrás del usuario, lo que da como resultado una percepción del sonido muy molesta. En particular en el caso en el que material de audio y vídeo se ofrece simultáneamente, como por ejemplo en el cine, una impresión de este tipo es especialmente molesta.

Una condición previa principal para la señal acústica de una película de cine (una banda sonora) consiste en que la impresión auditiva debe ser conforme a la impresión generada por las imágenes. Por tanto, las indicaciones audibles para la localización no deberían ir en contra de indicaciones visuales para la localización. Como consecuencia la voz correspondiente, cuando se ve a un hablante en la pantalla, también debería estar situada delante del usuario.

Lo mismo se aplica para todas las demás señales de audio, es decir no se limita obligatoriamente a situaciones en las que se ofrecen simultáneamente señales de audio y vídeo. Estas otras señales de audio son por ejemplo señales radiofónicas o audiolibros. Un oyente está acostumbrado a que la voz se genere desde los canales frontales, dándose probablemente la vuelta si de repente la voz procediese de los canales traseros para recuperar su impresión habitual.

Para mejorar la calidad de las señales de entorno se propone en la solicitud de patente alemana DE 102006017280.9-55 someter una señal de entorno extraída una vez a una detección de transitorios y realizar una supresión de transitorios sin llegar a pérdidas esenciales de energía en la señal de entorno. Para ello se realiza una sustitución de señales, para sustituir zonas con transitorios por señales correspondientes sin transitorios, aunque casi con la misma energía.

El documento de la AES Convention “Descriptor-based specialization”, J. Monceaux, F. Pachet, entre otros, 28 a 31 de mayo de 2005, Barcelona, España, da a conocer una espacialización basada en descriptores en la que, basándose en descriptores extraídos, se atenúa la voz detectada pasando a modo de silencio sólo el canal central. Para ello se utiliza un extractor de voz. Se utilizan un tiempo de inicio y un tiempo de estabilización para alisar modificaciones de la señal de salida. Así puede extraerse una banda sonora multicanal sin voz de una película. Cuando existe una determinada propiedad de reverberación estéreo en la señal de mezcla descendente estéreo original, esto lleva a que una herramienta de mezcla ascendente distribuya esta reverberación por todos los canales a excepción del canal central, de modo que puede oírse una reverberación. Para impedir esto se realiza un control de nivel dinámico para L, R, Ls y Rs para atenuar la reverberación de una voz.

El documento US 6 914 988 muestra por ejemplo un dispositivo para mejorar la reproducción de voz en un sistema multicanal. El objetivo de la presente invención consiste en crear un concepto para generar una señal multicanal con un número de canales de salida, que por un lado proporcione un producto flexible y por otro lado un producto de gran calidad.

Este objetivo se soluciona mediante un dispositivo para generar una señal multicanal según la reivindicación 1, un procedimiento para generar una señal multicanal según la reivindicación 23 o un programa informático según la reivindicación 24.

La presente invención se basa en el conocimiento de que se suprimen las componentes de voz en los canales traseros, esto es, en los canales de entorno, para que los canales traseros estén libres de componentes de voz. Para ello se mezcla de manera ascendente una señal de entrada con uno o varios canales para proporcionar un canal de señal directo y para proporcionar un canal de señal de entorno o, según la implementación, ya el canal de señal de entorno modificado. Está previsto un detector de voz para buscar componentes de voz en la señal de entrada, el canal directo o el canal de entorno, pudiendo aparecer por ejemplo tales componentes de voz en segmentos temporales y/o de frecuencia o también en elementos constitutivos de una descomposición ortogonal. Está previsto un modificador de señal para modificar la señal directa generada por el mezclador ascendente o una copia de la señal de entrada en el sentido de que, en ésta, se suprimen las componentes de señal de voz, mientras que las componentes de señal directa en los segmentos correspondientes que comprenden componentes de señal de voz, no se atenúan o se atenúan en

menor medida. Una señal de canal de entorno modificada de este tipo se utiliza entonces para la generación de señales de altavoz para altavoces correspondientes.

Sin embargo, si se ha modificado la señal de entrada, se utiliza directamente la señal de entorno generada por el mezclador ascendente, ya que en ésta ya se han suprimido las componentes de voz, ya que la señal de audio tomada como base ya tenía también componentes de voz suprimidas. Sin embargo, en este caso, cuando el proceso de mezcla ascendente también genera un canal directo, el canal directo no se calcula basándose en la señal de entrada modificada, sino basándose en la señal de entrada sin modificar para conseguir que las componentes de voz se supriman de manera selectiva y concretamente sólo en el canal de entorno, pero no en el canal directo, en el que las componentes de voz se desean de manera expresa.

De este modo se evita que en los canales traseros o canales de señal de entorno tenga lugar una reproducción de componentes de voz, que de lo contrario molestarían o incluso confundirían al oyente. Como consecuencia se garantiza según la invención que los diálogos y otro tipo de voz, inteligible para el oyente, que tiene por tanto una característica espectral típica para voz, se sitúen delante del oyente.

Los mismos requisitos existen también para el concepto en banda, en el que también se desea que las señales directas no se sitúen en los canales traseros, sino delante del oyente y dado el caso lateralmente respecto al oyente, pero no detrás del oyente, tal como se muestra en la figura 5c, en la que las componentes de señal directa (y también las componentes de señal de entorno) se sitúan delante del oyente.

Según la invención se realiza por tanto un procesamiento en función de la señal para eliminar o suprimir las componentes de voz en los canales traseros o en la señal de entorno. Para ello se realizas dos etapas esenciales, concretamente la detección de la aparición de voz y la supresión de voz, pudiendo realizarse la detección de la aparición de voz en la señal de entrada, en el canal directo o en el canal de entorno, y pudiendo realizarse la supresión de voz en el canal de entorno directa o indirectamente en la señal de entrada, que luego se utiliza para generar el canal de entorno, no utilizándose esta señal de entrada modificada para generar el canal directo.

Según la invención se consigue por tanto que, cuando se genera una señal envolvente multicanal a partir de una señal de audio con menos canales, que contiene componentes de voz, se garantice que las señales resultantes para los canales traseros desde el punto de vista del usuario comprendan una cantidad mínima de voz, con el fin de obtener la imagen acústica original delante del usuario (*front-image*). Si se reprodujera una cantidad especial de componentes de voz a través de los canales traseros, la posición de los hablantes se situaría fuera de la zona frontal, y concretamente en algún lugar entre el oyente y los altavoces frontales o en casos extremos incluso detrás del oyente. Esto daría como resultado una percepción del sonido muy molesta, en particular cuando las señales de audio se ofrecen simultáneamente con señales visuales, tal como es el caso por ejemplo en películas. Por ello muchas bandas sonoras de películas multicanal prácticamente no contienen componentes de voz en los canales traseros. Según la invención se detectan componentes de señal de voz y se suprimen en el lugar adecuado.

Ejemplos de realización preferidos de la presente invención se explican a continuación de manera detallada haciendo referencia a los dibujos adjuntos. Muestran:

la figura 1 un diagrama de bloques de un ejemplo de realización de la presente invención;

la figura 2 una asociación de segmentos de tiempo/frecuencia de una señal de análisis y un canal de entorno o señal de entrada para explicar los “segmentos correspondientes”;

la figura 3 una modificación de señal de entorno según un ejemplo de realización preferido de la presente invención;

la figura 4 una cooperación entre un detector de voz y un modificador de señal de entorno según otro ejemplo de realización de la presente invención;

la figura 5a un escenario de reproducción estéreo con fuentes directas (instrumentos de percusión) y componentes difusas;

la figura 5b un escenario de reproducción multicanal, en el que todas las fuentes directas se reproducen a través de los canales frontales y se reproducen componentes difusas a través de todos los canales, denominándose también este escenario concepto directo/de entorno;

la figura 5c un escenario de reproducción multicanal, en el que pueden reproducirse fuentes de sonido discretas también a través de canales traseros al menos parcialmente y en el que los canales de entorno no se reproducen o se reproducen en menor medida que en la figura 5b a través de los altavoces traseros;

la figura 6a otro ejemplo de realización con una detección de voz en el canal de entorno y una modificación del canal de entorno;

la figura 6b un ejemplo de realización con detección de voz en la señal de entrada y modificación del canal de entorno;

la figura 6c un ejemplo de realización con una detección de voz en la señal de entrada y una modificación de la señal de entrada;

la figura 6d otro ejemplo de realización con una detección de voz en la señal de entrada y una modificación en la señal de entorno, estando adaptada especialmente la modificación a la voz;

la figura 7 un ejemplo de realización con cálculo de factor de amplificación por bandas basándose en una señal paso banda/señal de subbanda; y

la figura 8 una representación detallada de un bloque de cálculo de amplificación de la figura 7.

La figura 1 muestra un diagrama de bloques de un dispositivo para generar una señal 10 multicanal, que en la figura 1 se muestra de modo que presenta un canal izquierdo L, un canal derecho R, un canal central C, un canal LFE, un canal trasero izquierdo LS y un canal trasero derecho RS. Ha de indicarse, sin embargo, que la presente invención también es adecuada para cualquier otra representación diferente de la representación 5.1 seleccionada para la misma, por ejemplo para una representación 7.1 o también para una representación 3.0, generándose en este caso sólo un canal izquierdo, un canal derecho y un canal central. La señal 10 multicanal con por ejemplo los seis canales mostrados en la figura 1, se genera a partir de una señal 12 de entrada o "x", que tiene un número de canales de entrada, siendo el número de canales de entrada 1 o mayor que 1 y por ejemplo igual a 2, cuando se da una mezcla descendente estéreo. Sin embargo, en general, el número de canales de salida es mayor que el número de canales de entrada.

El dispositivo mostrado en la figura 1 comprende un mezclador 14 ascendente para mezclar de manera ascendente la señal 12 de entrada, con el fin de generar al menos un canal 15 de señal directo y un canal 16 de señal de entorno o dado el caso un canal 16' de señal de entorno modificado. Además está previsto un detector 18 de voz que está configurado para utilizar como señal de análisis la señal 12 de entrada, tal como está previsto en 18a, o para utilizar el canal 15 de señal directo, tal como está previsto en 18b, o para utilizar otra señal que, con respecto a la aparición temporal/en frecuencia o con respecto a su característica relativa a las componentes de voz, sea similar a la señal 12 de entrada. El detector de voz detecta un segmento de la señal de entrada, del canal directo o por ejemplo también del canal de entorno, tal como se representa en 18c, en el que aparece una parte de voz. Esta parte de voz puede ser una parte de voz significativa, esto es, por ejemplo una parte de voz, cuya propiedad de voz se ha derivado en función de una medida cuantitativa o cualitativa determinada, superando la medida cualitativa y la medida cuantitativa un umbral, que también se denomina umbral de detección de voz.

En una medida cuantitativa se cuantifica una propiedad de voz con un valor numérico, y este valor numérico se compara con un umbral. En una medida cualitativa se toma una decisión por cada segmento, que puede tomarse según uno o varios criterios de decisión. Tales criterios de decisión pueden ser

por ejemplo diferentes características cuantitativas, que se comparan/ponderan entre sí o se procesan de algún modo para llegar a una decisión de sí/no.

El dispositivo mostrado en la figura 1 comprende además un modificador 20 de señal, que está configurado para modificar la señal de entrada original, tal como se muestra con 20a, o que está configurado para modificar el canal 16 de entorno. Cuando se modifica el canal 16 de entorno, el modificador 20 de señal emite un canal 21 modificado de entorno, mientras que cuando se modifica la señal 20a de entrada, se emite una señal 20b de entrada modificada hacia el mezclador 14 ascendente, que entonces genera el canal 16' de entorno modificado por ejemplo mediante la misma operación de mezcla ascendente utilizada para el canal 15 directo. En caso de que este proceso de mezcla ascendente también lleve, debido a la señal 20b de entrada modificada, a un canal directo, entonces este canal directo se rechazaría, ya que como canal directo se utiliza según la invención un canal directo derivado de la señal 12 de entrada sin modificar (sin supresión de voz) y no de la señal 20b de entrada modificada.

El modificador de señal está configurado para modificar segmentos del al menos un canal de entorno o de la señal de entrada, pudiendo ser por ejemplo estos segmentos, segmentos temporales o de frecuencia o partes de una descomposición ortogonal. En particular se modifican los segmentos que corresponden a los segmentos detectados por el detector de voz, de modo que el modificador de señal, tal como se ha representado, genera el canal 21 de entorno modificado o la señal 20b de entrada modificada, donde está atenuada o eliminada una parte de voz, estando la parte de voz en el segmento del canal directo correspondiente atenuada en menor medida o, mejor, no atenuada en absoluto.

Además el dispositivo mostrado en la figura 1 comprende un medio 22 de emisión de señal de altavoz para emitir señales de altavoz en un escenario de reproducción, como por ejemplo el escenario 5.1 mostrado a modo de ejemplo en la figura 1, siendo también posible sin embargo un escenario 7.1, un escenario 3.0 u otro escenario o escenario aún superior. En particular para la generación de las señales de altavoz para un escenario de reproducción se utiliza el al menos un canal directo y el al menos un canal de entorno modificado, pudiendo proceder el canal de entorno modificado o bien del modificador 20 de señal, tal como se muestra con 21, o del mezclador 14 ascendente, tal como se muestra con 16'.

Cuando por ejemplo se proporcionan dos canales 21 de entorno modificados, éstos dos canales de entorno modificados podrían alimentarse directamente a las dos señales de altavoz Ls, Rs, mientras que los canales directos sólo se alimentan a los tres altavoces frontales L, R, C, de modo que ha tenido lugar una división completa entre componentes

de señal de entorno y componentes de señal directa. Las componentes de señal directa se encuentran entonces todas delante del usuario y las componentes de señal de entorno se encuentran todas detrás del usuario. Como alternativa las componentes de señal de entorno pueden introducirse normalmente en un porcentaje menor también en los canales frontales, de modo que por ejemplo se produzca el escenario directo/de entorno mostrado en la figura 5b, en el que no sólo se generan señales de entorno por canales envolventes, sino también por los altavoces frontales por ejemplo L, C, R.

Por el contrario, si se prefiere el escenario en banda, entonces se emiten componentes de señal de entorno también principalmente por los altavoces frontales por ejemplo L, R, C, alimentándose sin embargo también componentes de señal directa al menos parcialmente a los dos altavoces traseros Ls, Rs. Concretamente para conseguir situar las dos fuentes 1100 y 1102 de señal directa en la figura 5c en los lugares mostrados, la parte de la fuente 1100 en el altavoz L será aproximadamente igual de grande que en el altavoz Ls, para que según una regla de panorámica típica la fuente 1100 pueda situarse en el centro entre L y Ls. El medio 22 de emisión de señal de altavoz puede provocar por tanto según la implementación un traspaso directo de un canal alimentado en el lado de entrada o puede realizar una correlación de los canales de entorno y de los canales directos, por ejemplo a través de un concepto en banda o un concepto directo/de entorno, de modo que tenga lugar una distribución de los canales por los altavoces individuales y, en última instancia, para generar la señal de altavoz real, puede producirse una suma de las partes a partir de los canales individuales.

La figura 2 muestra una división de tiempo/frecuencia de una señal de análisis en la sección superior y de un canal de entorno o señal de entrada en una sección inferior. En particular, a lo largo del eje horizontal está indicado el tiempo y a lo largo del eje vertical está indicada la frecuencia. Esto significa que, en la figura 2, para cada señal están representadas 15 casillas de tiempo/frecuencia o segmentos de tiempo/frecuencia, que en la señal de análisis y en el canal de entorno/señal de entrada tienen el mismo número. Esto significa que el modificador 20 de señal, por ejemplo cuando el detector 18 de voz detecta en el segmento 22 una señal de voz, procesa el segmento del canal de entorno/señal de entrada de algún modo, por ejemplo lo atenúa, elimina por completo o sustituye por una señal de síntesis que no tiene ninguna propiedad de voz. Ha de indicarse que en la presente invención la división no tiene que ser tan selectiva como se muestra en la figura 2. En su lugar ya una detección temporal también puede proporcionar un efecto satisfactorio, detectándose entonces un segmento temporal determinado de la señal de análisis, por ejemplo del segundo 2 al segundo 2,1, como que contiene señal de voz, para entonces procesar el segmento del canal de entorno o de la señal de entrada también entre el segundo 2 y el 2,1, para conseguir una supresión de voz.

Alternativamente también puede realizarse una descomposición ortogonal, por ejemplo por medio de un análisis de componente principal, utilizándose entonces tanto en el canal de entorno o la señal de entrada como en la señal de análisis la misma descomposición de componentes. Entonces en el canal de entorno o la señal de entrada se atenúan o suprimen o eliminan por completo determinadas componentes, detectadas como componentes de voz en la señal de análisis. Por tanto según la implementación se detecta un segmento en la señal de análisis, no procesándose entonces este segmento obligatoriamente en la señal de análisis, sino dado el caso también en otra señal.

La figura 3 muestra una implementación de un detector de voz en cooperación con un modificador de canal de entorno, proporcionando el detector de voz sólo una información de tiempo, esto es, cuando se considera la figura 2, sólo identifica en banda ancha el primer, segundo, tercer, cuarto o quinto segmento temporal y comunica esta información al modificador 20 de canal de entorno a través de una línea 18d de control (figura 1). El detector 18 de voz y el modificador 20 de canal de entorno, que funcionan de manera síncrona o con almacenamiento intermedio, consiguen ambos que en la señal que va a modificarse, que puede ser por ejemplo la señal 12 o la señal 16, la señal de voz o la componente de voz esté atenuada, mientras que se garantiza que una atenuación de este tipo del segmento correspondiente no aparezca o aparezca sólo en menor medida en el canal directo. Según la implementación esto puede conseguirse porque el mezclador 14 ascendente funcione sin tener en cuenta las componentes de voz, como por ejemplo en un procedimiento de matriz o en otro procedimiento que no realiza ningún procesamiento de voz especial. La señal directa así obtenida se suministra entonces sin un procesamiento adicional al medio 22 de emisión, mientras que la señal de entorno se procesa en cuanto a una supresión de voz.

Alternativamente cuando el modificador de señal somete la señal de entrada a una supresión de voz, el mezclador 14 ascendente puede funcionar en cierto modo dos veces, para por un lado, basándose en la señal de entrada original, extraer la componente de canal directo, pero basándose en la señal 20b de entrada modificada extraer el canal 16' de entorno modificado. En este caso se ejecutaría dos veces el mismo algoritmo de mezcla ascendente, aunque utilizando una señal de entrada diferente en cada caso, estando atenuada en una señal de entrada la componente de voz y no estando atenuada en la otra señal de entrada la componente de voz.

Según la implementación, el modificador de canal de entorno tiene una funcionalidad de una atenuación de banda ancha o una funcionalidad de un filtrado paso alto, tal como se explica mejor a continuación.

A continuación, mediante las figuras 6a, 6b, 6c y 6d se explican diferentes implementaciones del dispositivo según la invención.

En la figura 6a se extrae la señal de entorno a de la señal de entrada x, siendo esta extracción una parte de la funcionalidad del mezclador 14 ascendente. La aparición de voz se detecta en la señal de entorno a. El resultado de

detección d se utiliza en el modificador 20 de canal de entorno, que calcula la señal 21 de entorno modificada, en la que están suprimidas partes de voz.

La figura 6b muestra una configuración diferente a la figura 6a en la medida en que se suministra al detector 18 de voz como señal 18a de análisis la señal de entrada y no la señal de entorno. En particular la señal de canal de entorno modificada a_s se calcula de manera similar a la configuración de la figura 6a, detectándose sin embargo la voz en la señal de entrada. Esto es así porque las componentes de voz en general pueden encontrarse más fácilmente en la señal de entrada x que en la señal de entorno a. Así mediante la configuración mostrada en la figura 6b puede conseguirse una mayor fiabilidad.

En la figura 6c se extrae la señal de entorno con modificación de voz a_s de una versión x_s de la señal de entrada, que ya se ha sometido a una supresión de señal de voz. Como las componentes de voz en x aparecen normalmente de manera más prominente que en una señal de entorno extraída, su supresión se realiza de manera más segura y eficaz que en la figura 6a. Es desventajoso en la configuración mostrada en la figura 6c en comparación con la configuración en la figura 6a que pueden ampliarse aún más posibles artefactos de la supresión de voz y el proceso de extracción de entorno en función del tipo del procedimiento de extracción. En cualquier caso, en la figura 6c, la funcionalidad del extractor 14 de canal de entorno sólo se utiliza para extraer el canal de entorno de la señal de audio modificada. Sin embargo, el canal directo no se extrae de la señal (20b) de audio modificada x_s, sino basándose en la señal (12) de entrada original.

En la configuración mostrada en la figura 6d, la señal de entorno a se extrae de la señal de entrada x a través del mezclador ascendente. La aparición de voz se detecta en la señal de entrada x. Además a través de un analizador 30 de voz se calcula información secundaria adicional e, que controla adicionalmente la funcionalidad del modificador 20 de canal de entorno. Esta información secundaria se calcula directamente a partir de la señal de entrada y puede ser la posición de componentes de voz en una representación de tiempo/frecuencia, por ejemplo en forma de un espectrograma de la figura 2 o puede ser otro tipo de información adicional, a la que a continuación se hará referencia con más detalle.

A continuación se hace referencia con más detalle a la funcionalidad del detector 18 de voz. El objetivo de una detección de voz consiste en analizar una mezcla de señales de audio para estimar una probabilidad de que esté presente voz. La señal de entrada puede ser una señal que puede estar compuesta por una pluralidad de diferentes tipos de señales de audio, por ejemplo una señal de música, ruidos o efectos acústicos especiales, tal como se conocen en las películas de cine. Una posibilidad para la detección de voz consiste en utilizar un sistema de reconocimiento de patrones. Por reconocimiento de patrones se entiende el análisis de datos sin procesar y la realización de un procesamiento especial basándose en una categoría de un patrón que se ha descubierto en los datos sin procesar. En particular la expresión “patrón” o “*pattern*” describe una similitud básica que puede encontrarse entre las mediciones de objetos de categorías (clases) iguales. Las operaciones básicas de un sistema de reconocimiento de patrones consisten en la detección, es decir, la captación de los datos utilizando un convertidor, un procesamiento previo, una extracción de características y una clasificación, pudiendo realizarse estas operaciones básicas en el orden indicado.

Habitualmente se utilizan micrófonos como sensores para un sistema de reconocimiento de voz. Una preparación puede comprender una conversión A/D, un remuestreo o una reducción de ruido. La extracción de características es el cálculo de rasgos característicos para cada objeto a partir de las mediciones. Las características se seleccionan de modo que sean similares entre objetos de la misma clase, consiguiéndose por tanto una buena compacidad dentro de las clases y de modo que sean diferentes para objetos de diferentes clases, de modo que se consigue una separabilidad entre clases. Un tercer requisito consiste en que las características deben ser robustas respecto al ruido, las condiciones del entorno y transformaciones de la señal de entrada irrelevantes para la percepción por el ser humano. La extracción de características puede dividirse en dos fases separadas. La primera fase es el cálculo de características y la segunda fase es la proyección de características o transformación sobre una base en general ortogonal, para minimizar una correlación entre vectores de características y reducir la dimensionalidad de las características, no utilizándose elementos con baja energía.

La clasificación es el proceso de la decisión de si existe voz o no, y concretamente basándose en las características extraídas y un clasificador entrenado. Así se da la siguiente ecuación.

$$\Omega_{XY} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}, \mathbf{x}_i \in \Re^n, y \in Y = \{1, \ldots, c\}$$

En la ecuación anterior se define una cantidad de vectores de entrenamiento Ω_{xy}, designándose los vectores de características por x_i y el conjunto de clases por Y. Para una detección de voz básica se aplica por tanto que Y tiene dos valores, concretamente {voz, no voz}.

En la fase de entrenamiento se calculan las características x_i a partir de datos especificados, es decir a partir de señales de audio en las que se sabe a qué clases y pertenecen. Tras una finalización del entrenamiento el clasificador ha aprendido las características de todas las clases.

En la fase de aplicación del clasificador se calculan y proyectan las características a partir de los datos desconocidos igual que en la fase de entrenamiento y el clasificador, gracias al conocimiento adquirido en el entrenamiento, las

clasifica a través de las características de las clases.

A continuación se hace referencia a implementaciones especiales de la supresión de voz, tal como pueden realizarse por ejemplo a través del modificador 20 de señal. Así pueden utilizarse diferentes procedimientos para suprimir voz en una señal de audio. En este caso hay procedimientos que se conocen a partir del campo de la amplificación de voz y la reducción de ruido para aplicaciones de comunicación. Originalmente los procedimientos de amplificación de voz se utilizaron para amplificar la voz en una mezcla de voz y ruidos de fondo. Tales métodos pueden modificarse para producir también lo contrario, concretamente una supresión de voz, tal como se realiza para la presente invención.

Así, existen enfoques para la amplificación de voz y la reducción de ruido que atenúan o amplifican los coeficientes de una representación de tiempo/frecuencia según un valor de estimación del grado del ruido contenido en un coeficiente de tiempo/frecuencia. Cuando no se conoce información adicional sobre un ruido de fondo, por ejemplo información a priori o información medida mediante un sensor de ruido especial, se obtiene una representación de tiempo/frecuencia a partir de una medición ruidosa, por ejemplo utilizando procedimientos estadísticos mínimos especiales. Una regla de supresión de ruido calcula un factor de atenuación utilizando el valor de estimación de ruido. Este principio se conoce como atenuación espectral a corto plazo o ponderación espectral, tal como se conoce por ejemplo en G. Schmid, "Single-channel noise suppression based on spectral weighting", Eurasip Newsletter 2004. Los métodos de procesamiento de señales que funcionan según el principio de la atenuación espectral a corto plazo (STSA) consisten en la sustracción espectral, el filtrado de Wiener y el algoritmo de Ephraim-Malah. Una formulación más general del enfoque de STSA lleva a un procedimiento de subespacio de señales que también se conoce como método del rango reducido y se describe en P. Hansen y S. Jensen, "Fir filter representation of reduced-rank noise reduction", IEEE TSP, 1998.

En principio, por tanto, pueden utilizarse todos los procedimientos que amplifiquen la voz o supriman componentes que no son de voz, de manera opuesta a su uso conocido, para suprimir voz o amplificar elementos que no sean de voz. El modelo general de la amplificación de voz o supresión de ruido consiste en que la señal de entrada es una mezcla de una señal (voz) deseada y el ruido de fondo (no voz). Una supresión de la voz se consigue por ejemplo mediante la inversión de los factores de atenuación en un procedimiento basado en STSA o cambiando la definición de la señal deseada y el ruido de fondo.

Un requisito importante en la supresión de voz consiste sin embargo en que, en cuanto al contexto de la mezcla ascendente, la señal de audio resultante se percibe como señal de audio de alta calidad de audio. Se conoce que los procedimientos de mejora de voz y los procedimientos de reducción de ruido introducen artefactos audibles en la señal de salida. Un ejemplo de un artefacto de este tipo se conoce como ruido musical o sonidos musicales y resulta de una estimación errónea de umbrales mínimos de ruido (*noise floors*) y factores de atenuación de subbanda oscilantes.

Alternativamente pueden utilizarse también procedimientos de separación de fuente ciegos para separar las partes de señal de voz de la señal de entorno y manipular ambas a continuación por separado.

Para el requisito especial de la generación de señales de audio de alta calidad se prefieren sin embargo determinados procedimientos explicados a continuación debido al hecho de que en comparación con otros procedimientos tienen esencialmente un mejor rendimiento. Un procedimiento consiste en la atenuación de banda ancha, tal como se indica con 20 en la figura 3. La señal de audio se atenúa en los segmentos temporales en los que existe voz. Los factores de amplificación especiales se encuentran en el intervalo entre -12 dB y -3 dB, situándose una atenuación preferida en 6 dB. Puesto que del mismo modo se suprimen otras componentes/partes de señal, podría pensarse que toda la pérdida de energía de señal de audio se percibe de manera clara. Sin embargo se ha demostrado que este efecto no es molesto ya que el usuario se concentra sin más en particular en los altavoces frontales L, C, R, cuando comienza una secuencia de voz, de modo que el usuario no percibirá la reducción de energía de los canales traseros o la señal de entorno cuando se concentre precisamente en una señal de voz. Esto se ve reforzado en particular por el efecto típico adicional de que el nivel de la señal de audio aumenta sin más debido a una voz introducida. Mediante la introducción de una atenuación en el intervalo entre -12 dB y 3 dB no se percibe la atenuación como molesta. En su lugar el usuario percibe de manera mucho más agradable que, gracias a la supresión de componentes de voz en los canales traseros, se consigue un efecto que lleva a que para el usuario las componentes de voz estén situadas exclusivamente en los canales frontales.

Un procedimiento alternativo que también se indica en 20 en la figura 3, consiste en un filtrado paso alto. La señal de audio se somete a un filtrado paso alto donde existe voz, situándose una frecuencia límite en el intervalo entre 600 Hz y 3.000 Hz. El ajuste de la frecuencia límite se obtiene a partir de la característica de señal de voz en cuanto a la presente invención. El espectro de potencia a largo plazo de una señal de voz se concentra en un intervalo por debajo de los 2,5 kHz. El intervalo preferido de la frecuencia fundamental de voz tonal (*voiced speech*) se encuentra en el intervalo entre 75 Hz y 330 Hz. Un intervalo entre 60 Hz y 250 Hz se obtiene para adultos hombres. Los valores medios se encuentran en 120 Hz para hablantes hombres y 215 para hablantes mujeres. Debido a las resonancias en el tracto vocal se amplifican determinadas frecuencias de señal. Los picos correspondientes en el espectro se denominan también frecuencias de formante o simplemente formantes. Normalmente existen aproximadamente tres formantes significativos por debajo de los 3.500 Hz. Como consecuencia la voz muestra una naturaleza 1/F, es decir la energía espectral disminuye a medida que aumenta la frecuencia. Por tanto las componentes de voz pueden filtrarse bien, para los fines de la presente invención, mediante un filtrado paso alto con el intervalo de frecuencia límite indicado.

Una implementación preferida adicional consiste en el modelado de señal sinusoidal que se representa mediante la figura 4. Así en una primera etapa 40 se detecta la onda fundamental de una voz, pudiendo tener lugar esta detección en el detector 18 de voz aunque también, tal como se muestra en la figura 6e, en el analizador 30 de voz. A continuación en una etapa 41 se realiza un examen para encontrar los armónicos pertenecientes a la onda fundamental. Esta funcionalidad puede realizarse en el detector de voz/analizador de voz o incluso ya en el modificador de señal de entorno. A continuación se calcula para la señal de entorno un espectrograma, y concretamente basándose en una transformación directa realizada por bloques, tal como se explica en 42. A continuación se realiza la verdadera supresión de voz en una etapa 43, en la que se atenúan la onda fundamental y los armónicos en el espectrograma. En una etapa 44 se somete entonces la señal de entorno modificada, en la que se han atenuado o eliminado la onda fundamental y los armónicos, de nuevo a una transformación inversa para conseguir la señal de entorno modificada o la señal de entrada modificada.

Este modelado de señal sinusoidal se utiliza a menudo para la síntesis de sonidos, la codificación de audio, la separación de fuentes, la manipulación de sonidos y la supresión de ruido. En este caso se representa una señal como composición de ondas sinusoidales con frecuencias y amplitudes variables en el tiempo. Las componentes de señal de voz tonales se manipulan identificándose y modificándose los sonidos parciales, es decir la onda fundamental y sus armónicos.

Los sonidos parciales se identifican por medio de un localizador de sonidos parciales tal como se muestra con 41. Normalmente la localización de sonido parcial se realiza en los dominios de tiempo/frecuencia. Se realiza un espectrograma por medio de una transformada de Fourier a corto plazo tal como se indica con 42. Se detectan máximos locales en cada espectro del espectrograma y se determinan trayectorias a través de máximos locales de espectros adyacentes. Una estimación de la frecuencia fundamental puede ayudar al proceso de selección de picos (*peak picking*), realizándose esta estimación de la frecuencia fundamental en 40. Una representación de señal sinusoidal se consigue entonces a partir de las trayectorias. Ha de indicarse que también puede variarse el orden entre la etapa 40, 41 y la etapa 42 de modo que en primer lugar se realice una transformación 42 directa, que tiene lugar en el analizador 30 de voz de la figura 6d.

Se han propuesto diferentes ampliaciones de la derivación de una representación de señal sinusoidal. En D. Andersen y M. Clements, “Audio signal noise reduction using multi-resolution sinusoidal modeling”, Proceedings of ICASSP 1999 se representa un enfoque de procesamiento multirresolución para la reducción de ruido. Un proceso iterativo para la derivación de la representación sinusoidal se presentó en J. Jensen y J. Hansen, “Speech enhancement using a constrained iterative sinusoidal model”, IEEE TSAP 2001.

Mediante el uso de la representación de señal sinusoidal se obtiene una señal de voz mejorada mediante la amplificación de la componente sinusoidal. La supresión de voz según la invención pretende conseguir sin embargo justo lo contrario, concretamente suprimir los sonidos parciales, comprendiendo los sonidos parciales la onda fundamental y sus armónicos, y concretamente para un segmento de voz con voz tonal. Normalmente las componentes de voz con alta energía son tonales. Así se emite una voz a un nivel de 60 - 75 dB para vocales y aproximadamente 20 - 30 dB menos para consonantes. Para voz tonal (vocales) la excitación es una señal pulsada periódica. La señal de excitación se filtra a través del tracto vocal. Por consiguiente casi toda la energía de un segmento de voz tonal se concentra en la onda fundamental y sus armónicos. Mediante la supresión de estos sonidos parciales se suprimen las componentes de voz de manera significativa.

Otra manera de conseguir una supresión de voz se representa en la figura 7 y la figura 8. La figura 7 y la figura 8 explican el principio fundamental de la atenuación espectral a corto plazo o ponderación espectral. En este caso se estima en primer lugar el espectro de densidad de potencia del ruido de fondo. El procedimiento representado estima la cantidad de voz contenida en una casilla de tiempo/frecuencia utilizando denominadas características de bajo nivel que indican una medida del “tipo de voz” de una señal en un segmento de frecuencia determinado. Las características o propiedades de bajo nivel del plano inferior son características con bajo nivel respecto a la interpretación de su significado y el esfuerzo de su cálculo.

La señal de audio se descompone en un número de bandas de frecuencia por medio de un banco de filtros o una transformada de Fourier a corto plazo que se representa en 70 en la figura 7. A continuación, tal como se representa a modo de ejemplo en 71a y 71b, se calculan factores de amplificación variables en el tiempo para todas las subbandas a partir de tales características de planos inferiores (*low-level-features*), para atenuar señales de subbanda de manera proporcional a la cantidad de voz que contienen. Características adecuadas en un plano inferior son la medida de planeidad espectral (SFM; SFM = *spectral flatness measure*) y la energía de modulación de 4 Hz (4HzME). La SFM mide el grado de tonalidad de una señal de audio y se obtiene para una banda a partir del cociente del valor medio geométrico de todos los valores espectrales en una banda y del valor medio aritmético de las componentes espectrales en la banda. La 4HzME viene dada porque la voz tiene un pico de modulación de energía característico a aproximadamente 4 Hz, lo que corresponde a la tasa de sílabas media de un hablante.

La figura 8 muestra una representación detallada del bloque 71a y 71b de cálculo de amplificación de la figura 7. Basándose en una subbanda x_i se calcula una pluralidad de diferentes característica de bajo nivel, esto es, LLF1, …, LLFn. Estas características se combinan entonces en un combinador 80, para dar un factor de amplificación g_i para una subbanda.

Ha de indicarse que, según la implementación, no tienen que utilizarse obligatoriamente características de bajo orden, sino que puede utilizarse cualquier característica, como por ejemplo también características de energía, etc., que entonces pueden combinarse entre sí según la implementación de la figura 8 en un combinador, para dar un factor de amplificación cuantitativo g_i, de modo que cada banda (en cualquier instante) se atenúa de manera variable, para conseguir una supresión de voz.

En función de las circunstancias el procedimiento según la invención puede implementarse en hardware o en software. La implementación puede producirse en un medio de almacenamiento digital, en particular un disquete o CD con señales de control legibles electrónicamente, que pueden actuar conjuntamente con un sistema informático programable de modo que se realiza el procedimiento. En general la invención consiste por tanto también en un producto de programa informático con un código de programa almacenado en un soporte legible por máquina para la realización del procedimiento según la invención, cuando el producto de programa informático se ejecuta en un ordenador. Dicho de otro modo la invención puede realizarse por tanto como un programa informático con un código de programa para la realización del procedimiento cuando el programa informático se ejecuta en un ordenador.

REIVINDICACIONES

1. Dispositivo para generar una señal (10) multicanal con un número de señales de canal de salida, que es mayor que un número de señales de canal de entrada de una señal (12) de entrada, siendo el número de señales de canal de entrada igual a 1 o mayor, con las características siguientes:

un mezclador (14) ascendente para mezclar de manera ascendente la señal de entrada, que presenta una parte de voz, para proporcionar al menos una señal de canal directo y al menos una señal de canal de entorno con una parte de voz;

un detector (18) de voz para detectar un segmento de la señal de entrada, de la señal de canal directo o de la señal de canal de entorno, en el que aparece la parte de voz; y

un modificador (20) de señal para modificar un segmento de la señal de canal de entorno, que corresponde al segmento detectado por el detector (18), para obtener una señal de canal de entorno modificada, en la que la parte de voz está atenuada o eliminada, no estando atenuado o estando atenuado en menor medida el segmento en la señal de canal directo; y

un medio (22) de emisión de señal de altavoz para emitir señales de altavoz en un esquema de reproducción utilizando el canal directo y la señal de canal de entorno modificada, siendo las señales de altavoz las señales de canal de salida.

2. Dispositivo según la reivindicación 1, en el que el medio (22) de emisión de señal de altavoz está configurado para funcionar según un esquema directo/de entorno, en el que cada canal directo puede correlacionarse con un altavoz propio, y cada señal de canal de entorno puede correlacionarse con un altavoz propio, estando configurado el medio (22) de emisión de señal de altavoz para correlacionar con señales de altavoz, para altavoces situados detrás de un oyente en el esquema de reproducción, sólo la señal de canal de entorno y no el canal directo.

3. Dispositivo según la reivindicación 1, en el que el medio (22) de emisión de señal de altavoz está configurado para funcionar según un esquema en banda, en el que cada señal de canal directo puede correlacionarse con uno o varios altavoces en función de su posición, y en el que el medio (22) de emisión de señal de altavoz está configurado para sumar la señal de canal de entorno y el canal directo o una parte de la señal de canal de entorno o del canal directo, que están determinadas para un altavoz, con el fin de obtener una señal de emisión de altavoz para el altavoz.

4. Dispositivo según una de las reivindicaciones anteriores, en el que el medio de emisión de señal de altavoz está configurado para proporcionar señales de altavoz para al menos tres canales, que en el esquema de reproducción pueden situarse delante de un oyente, y para generar al menos dos canales, que en el esquema de reproducción pueden situarse detrás del oyente.

5. Dispositivo según una de las reivindicaciones anteriores,

en el que el detector (18) de voz está configurado para funcionar por bloques en el tiempo, y para analizar cada bloque temporal de manera selectiva en frecuencia por bandas, con el fin de detectar una banda de frecuencia para un bloque temporal, y

en el que el modificador (20) de señal está configurado para modificar una banda de frecuencia en un bloque temporal de la señal de canal de entorno que corresponde a la banda detectada por el detector (18) de voz.

6. Dispositivo según una de las reivindicaciones anteriores,

en el que el modificador de señal está configurado para atenuar la señal de canal de entorno o partes de la señal de canal de entorno en un intervalo de tiempo detectado por el detector (18) de voz, y

estando configurados el mezclador (14) ascendente y el medio (22) de emisión de señal de altavoz para generar el al menos un canal directo de manera que el mismo segmento temporal no se atenúe o se atenúe en menor medida, de modo que el canal directo presente una componente de voz, que en una reproducción pueda percibirse con más intensidad que una componente de voz en la señal de canal de entorno modificada.

7. Dispositivo según una de las reivindicaciones anteriores, en el que el modificador (20) de señal está configurado para someter la al menos una señal de canal de entorno a un filtrado paso alto cuando el detector (18) de voz ha detectado un segmento temporal en el que aparece una parte de voz, situándose una frecuencia límite del filtro paso alto entre 400 Hz y 3.500 Hz.

8. Dispositivo según una de las reivindicaciones anteriores,

en el que el detector (18) de voz está configurado para detectar una aparición temporal de una componente de señal de voz, y

en el que el modificador (20) de señal está configurado para determinar una frecuencia fundamental de la componente de señal de voz, y

para atenuar (43) selectivamente sonidos en la señal de canal de entorno o la señal de entrada a la frecuencia fundamental y los armónicos, con el fin obtener la señal de canal de entorno modificada o la señal de entrada modificada.

9. Dispositivo según una de las reivindicaciones anteriores,

en el que el detector (18) de voz está configurado para determinar por cada banda de frecuencia una medida de un contenido de voz, y

en el que el modificador (20) de señal está configurado para atenuar (72a, 72b) una banda correspondiente de la señal de canal de entorno según la medida con un factor de atenuación, resultando una medida superior en un factor de atenuación superior y una medida inferior en un factor de atenuación inferior.

10. Dispositivo según la reivindicación 9, en el que el modificador (20) de señal presenta las características siguientes:

un convertidor (70) de dominio de tiempo-frecuencia para convertir la señal de entorno en una representación espectral;

un atenuador (72a, 72b) para la atenuación variable de manera selectiva en frecuencia de la representación espectral; y

un convertidor (73) de dominio de frecuencia-tiempo para convertir la representación espectral atenuada de manera variable al dominio de tiempo, con el fin de obtener la señal de canal de entorno modificada.

11. Dispositivo según la reivindicación 9 ó 10, en el que el detector (18) de voz presenta las características siguientes:

un convertidor (42) de dominio de tiempo-frecuencia para proporcionar una representación espectral de una señal de análisis;

un medio para calcular una o varias características (71a, 71b) por cada banda de la señal de análisis; y

un medio (80) para calcular una medida de un contenido de voz basándose en una combinación de la una o varias características por cada banda.

12. Dispositivo según la reivindicación 11, en el que el modificador (20) de señal está configurado para calcular como características una medida de planeidad espectral (SFM) o una energía de modulación de 4 Hz (4HzME).

13. Dispositivo según una de las reivindicaciones anteriores, en el que el detector (18) de voz está configurado para analizar la señal (18c) de canal de entorno, y en el que el modificador (20) de señal está configurado para modificar la señal (16) de canal de entorno.

14. Dispositivo según una de las reivindicaciones 1 a 12, en el que el detector (18) de voz está configurado para analizar la señal (18a) de entrada, y en el que el modificador (20) de señal está configurado para modificar la señal (16) de canal de entorno basándose en información (18d) de control del detector (18) de voz.

15. Dispositivo según una de las reivindicaciones 1 a 12, en el que el detector (18) de voz está configurado para analizar la señal (18a) de entrada, y en el que el modificador (20) de señal está configurado para modificar la señal de entrada basándose en información (18d) de control del detector (18) de voz, y en el que el mezclador (14) ascendente presenta un extractor de canal de entorno, que está configurado para, basándose en la señal de entrada modificada, determinar la señal (16') de canal de entorno modificada, estando configurado además el mezclador (14) ascendente para, basándose en la señal (12) de entrada en la entrada del modificador (20) de señal, determinar la señal (15) de canal directo.

16. Dispositivo según una de las reivindicaciones 1 a 12,

en el que el detector (18) de voz está configurado para analizar la señal (18a) de entrada, en el que además está previsto un analizador (30) de voz, para someter la señal de entrada a un análisis de voz, y

en el que el modificador (20) de señal está configurado para modificar la señal (16) de canal de entorno basándose en información (18d) de control del detector (18) de voz y basándose en información (18e) de análisis de voz del analizador (30) de voz.

17. Dispositivo según una de las reivindicaciones anteriores, en el que el mezclador (14) ascendente está configurado como decodificador de matriz.

18. Dispositivo según una de las reivindicaciones anteriores, en el que el mezclador (14) ascendente está configurado como mezclador ascendente ciego que, basándose únicamente en la señal (12) de entrada, pero sin información de mezcla ascendente transmitida de manera adicional, genera la señal (15) de canal directo y la señal (16) de canal de entorno.

19. Dispositivo según una de las reivindicaciones anteriores,

en el que el mezclador (14) ascendente está configurado para realizar un análisis estadístico de la señal (12) de entrada, con el fin de generar la señal (15) de canal directo y la señal (16) de canal de entorno.

20. Dispositivo según una de las reivindicaciones anteriores, en el que la señal de entrada es una señal mono con un canal y en el que la señal de salida es una señal multicanal con dos o más señales de canal.

21. Dispositivo según una de las reivindicaciones 1 a 19, en el que el mezclador (14) ascendente está configurado para obtener como señal de entrada una señal estéreo con dos señales de canal estéreo, y en el que el mezclador (14) ascendente está configurado además para realizar la señal (16) de canal de entorno basándose en un cálculo de correlación cruzada de las señales de canal estéreo.

22. Procedimiento para generar una señal (10) multicanal con un número de señales de canal de salida, que es mayor que un número de señales de canal de entrada de una señal (12) de entrada, siendo el número de señales de canal de entrada igual a 1 o mayor, con las etapas siguientes:

mezclar de manera ascendente (14) la señal de entrada, para proporcionar al menos una señal de canal directo y al menos una señal de canal de entorno;

detectar (18) un segmento de la señal de entrada, de la señal de canal directo o de la señal de canal de entorno, en el que aparece una parte de voz; y

modificar (20) un segmento de la señal de canal de entorno que corresponde al segmento detectado en la etapa de la detección (18), para obtener una señal de canal de entorno modificada, en la que la parte de voz está atenuada o eliminada, no estando atenuado o estando atenuado en menor medida el segmento en la señal de canal directo; y

emitir (22) señales de altavoz en un esquema de reproducción utilizando el canal directo y la señal de canal de entorno modificada, siendo las señales de altavoz las señales de canal de salida.

23. Programa informático con un código de programa para realizar el procedimiento según la reivindicación 22, cuando el código de programa se ejecuta en un ordenador.

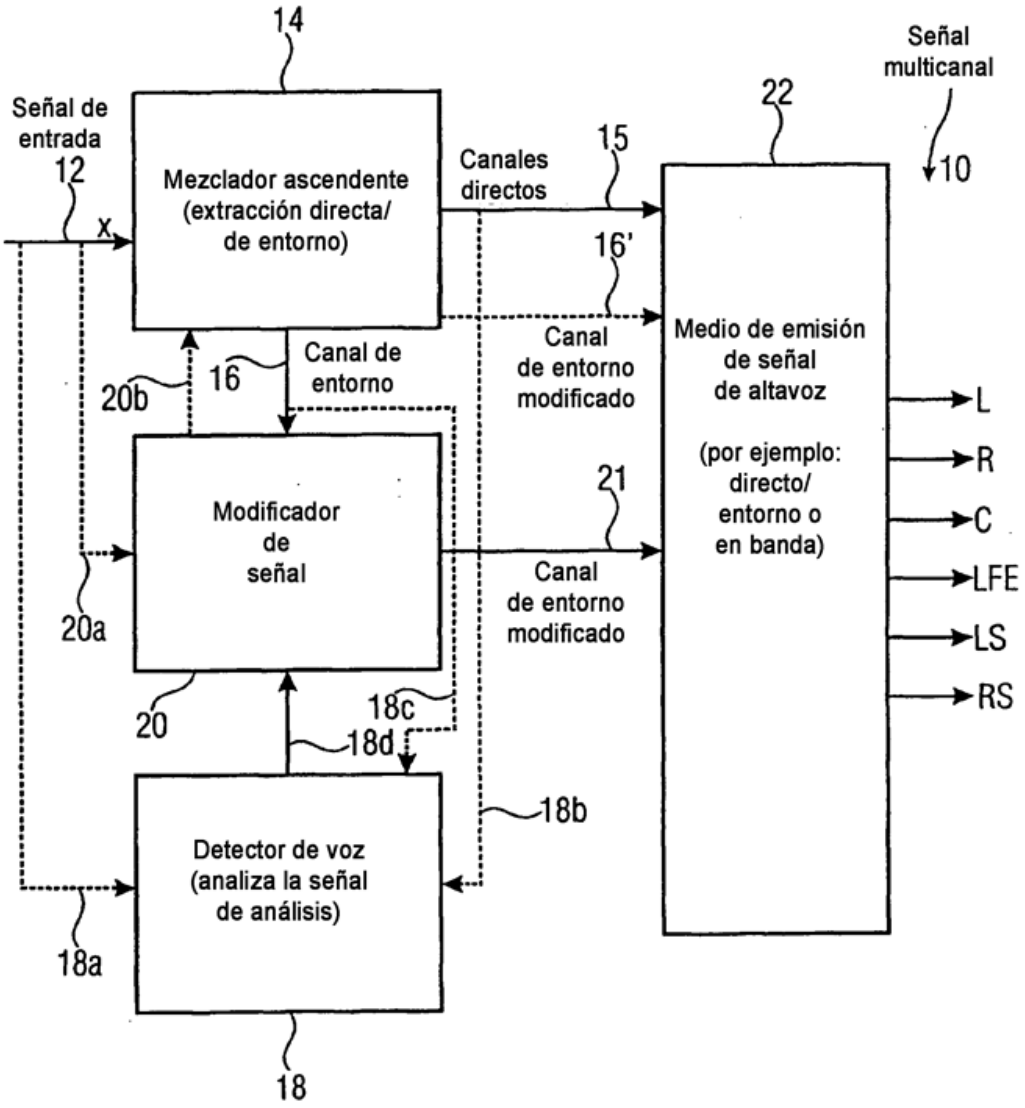
14
Señal de entrada
12
x
Mezclador ascendente (extracción directa/ de entorno)
Canales directos
15
22
Señal multicanal
10
16'
Canal de entorno modificado
Medio de emisión de señal de altavoz
(por ejemplo: directo/ entorno o en banda)
L
R
C
LFE
LS
RS
20b
16
Canal de entorno
Modificador de señal
21
Canal de entorno modificado
20a
20
18c
18d
18b
Detector de voz (analiza la señal de análisis)
18a
18

FIGURA 1

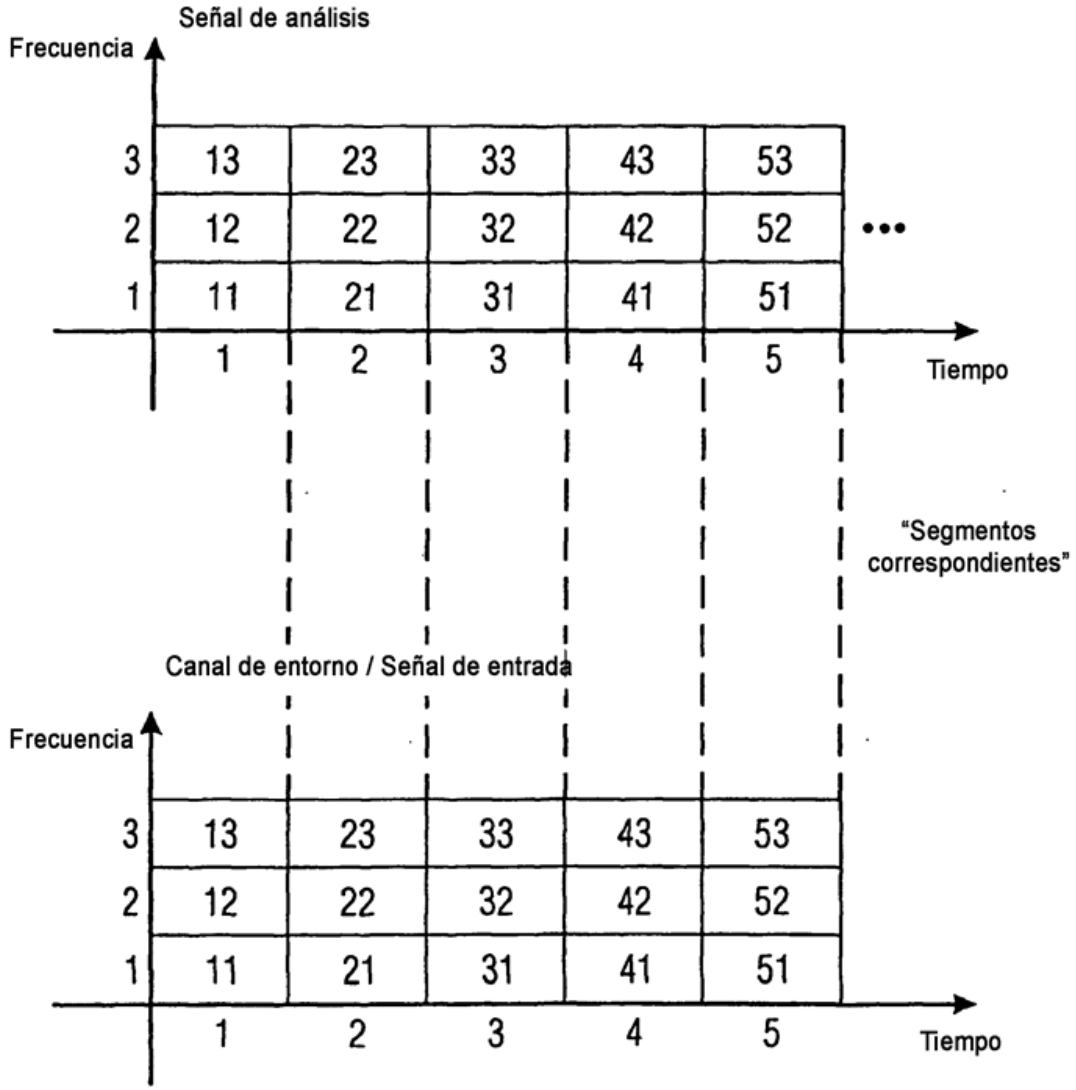
Señal de análisis
Frecuencia
3
2
1
13
23
33
43
53
12
22
32
42
52
11
21
31
41
51
1
2
3
4
5
Tiempo
"Segmentos correspondientes"
Canal de entorno / Señal de entrada
Frecuencia
3
2
1
13
23
33
43
53
12
22
32
42
52
11
21
31
41
51
1
2
3
4
5
Tiempo

FIGURA 2

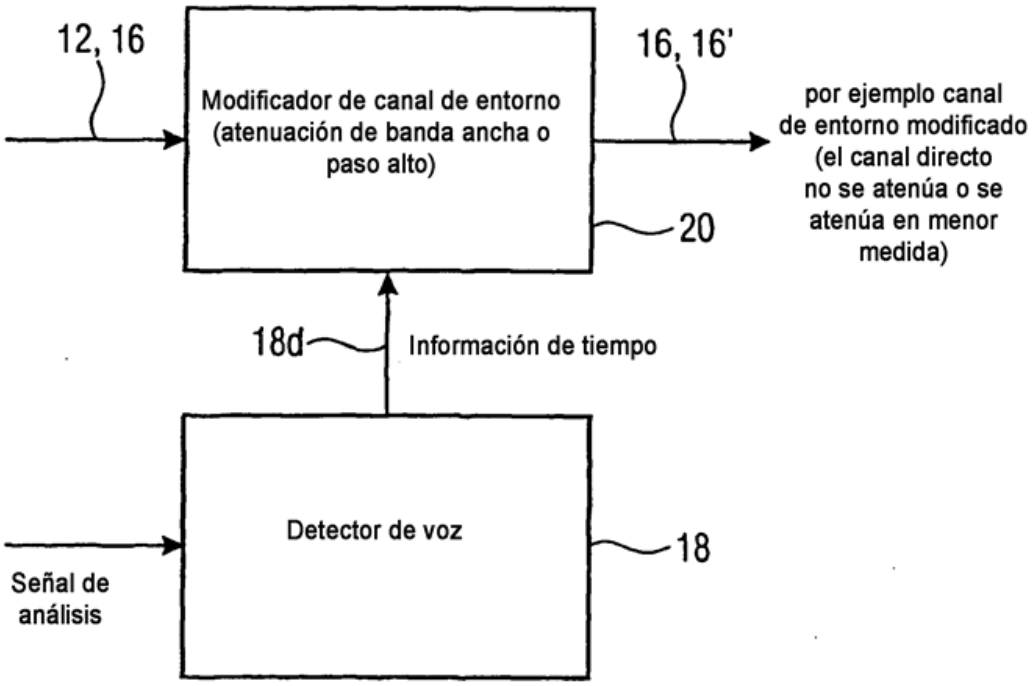
12, 16
Modificador de canal de entorno (atenuación de banda ancha o paso alto)
16, 16'
por ejemplo canal de entorno modificado (el canal directo no se atenúa o se atenúa en menor medida)
20
18d
Información de tiempo
Detector de voz
18
Señal de análisis

FIGURA 3

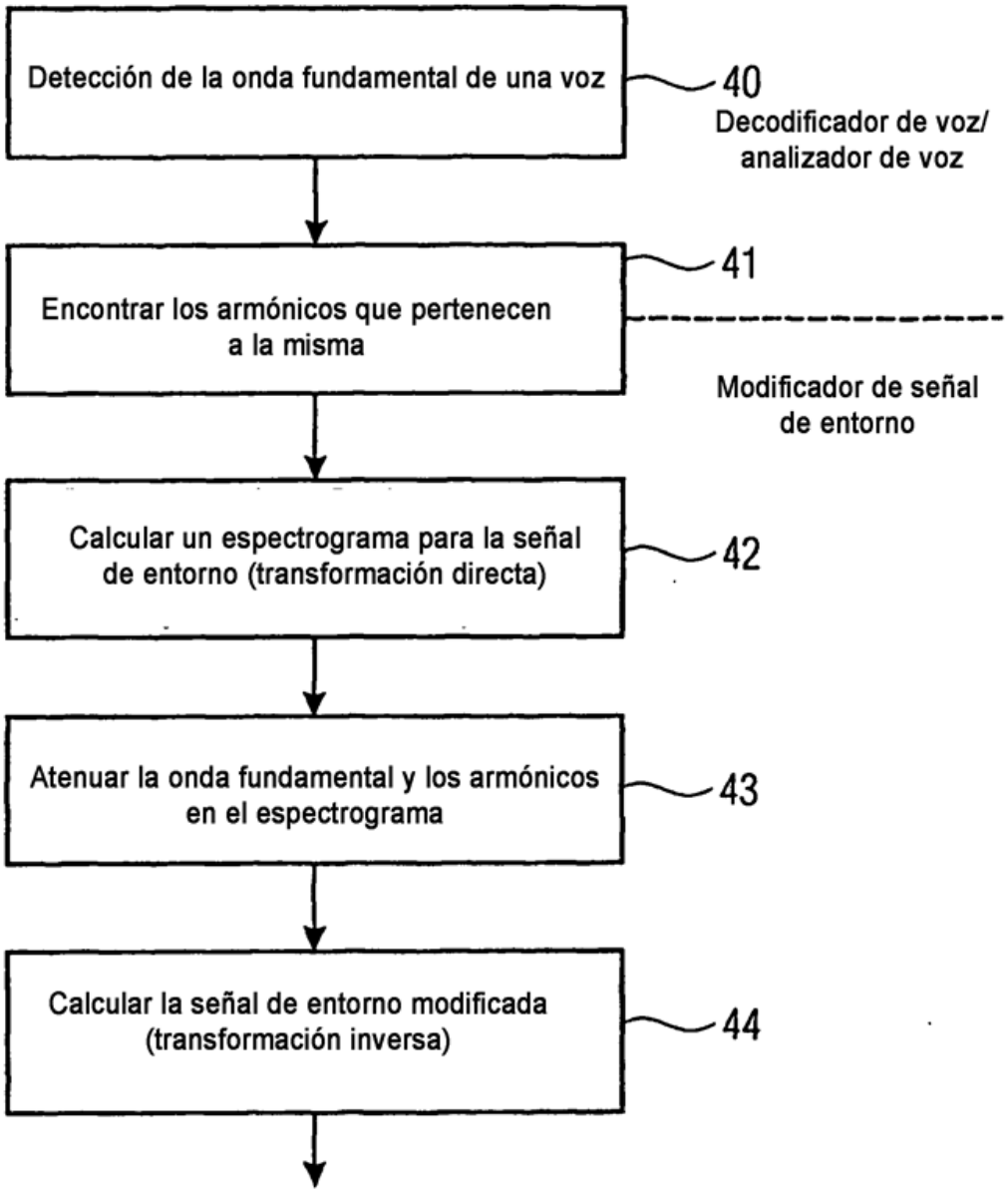
Detección de la onda fundamental de una voz
40
Decodificador de voz/ analizador de voz
Encontrar los armónicos que pertenecen a la misma
41
Modificador de señal de entorno
Calcular un espectrograma para la señal de entorno (transformación directa)
42
Atenuar la onda fundamental y los armónicos en el espectrograma
43
Calcular la señal de entorno modificada (transformación inversa)
44

FIGURA 4

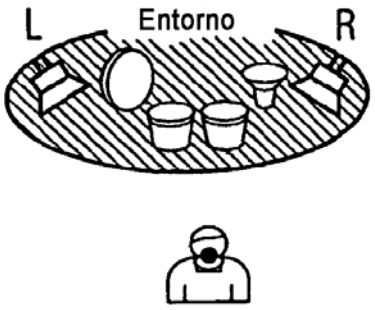

FIGURA 5A

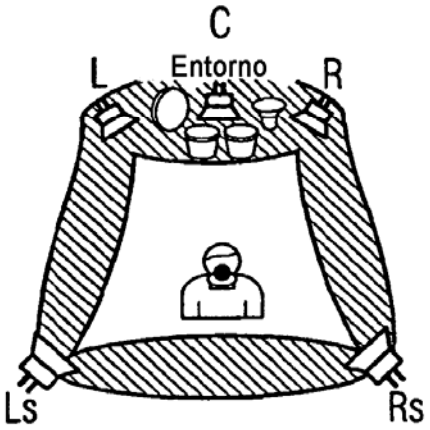

FIGURA 5B
(Directo/entorno)

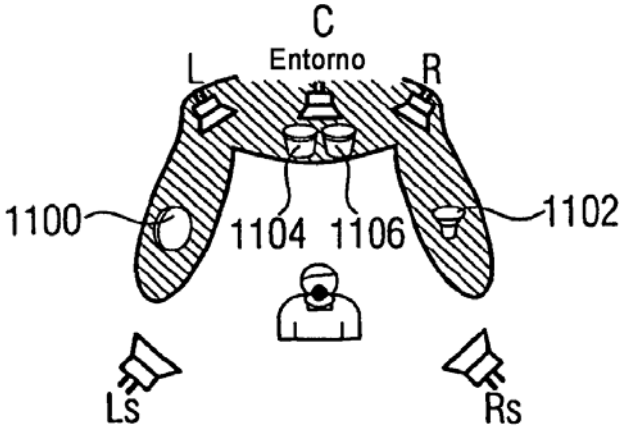

FIGURA 5C
(en banda)

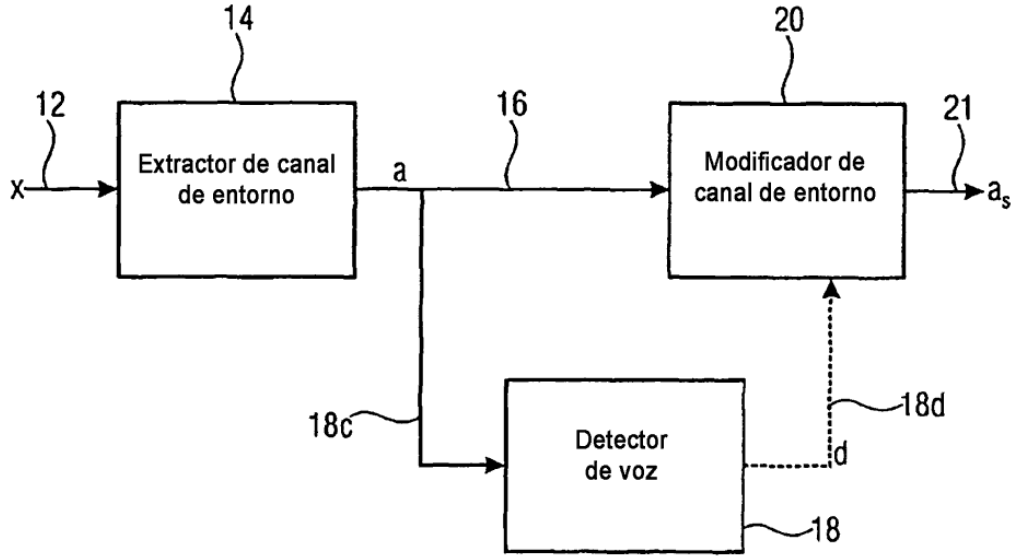

FIGURA 6A

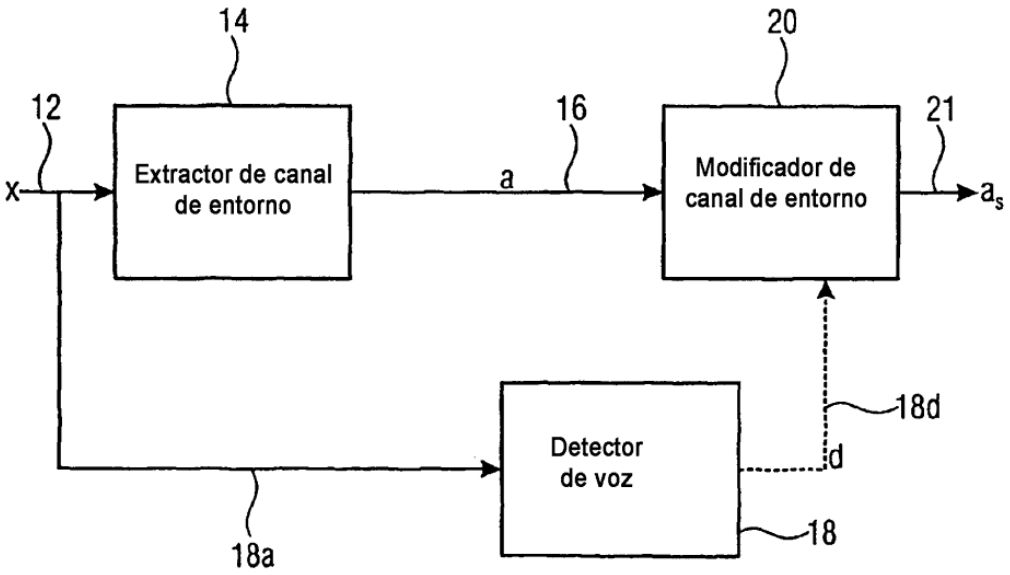

FIGURA 6B

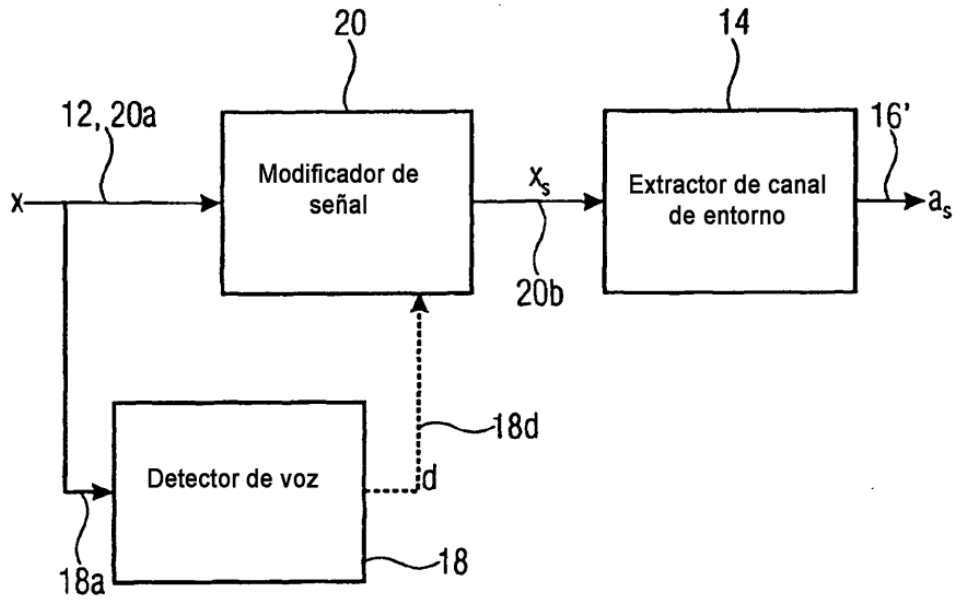

FIGURA 6C

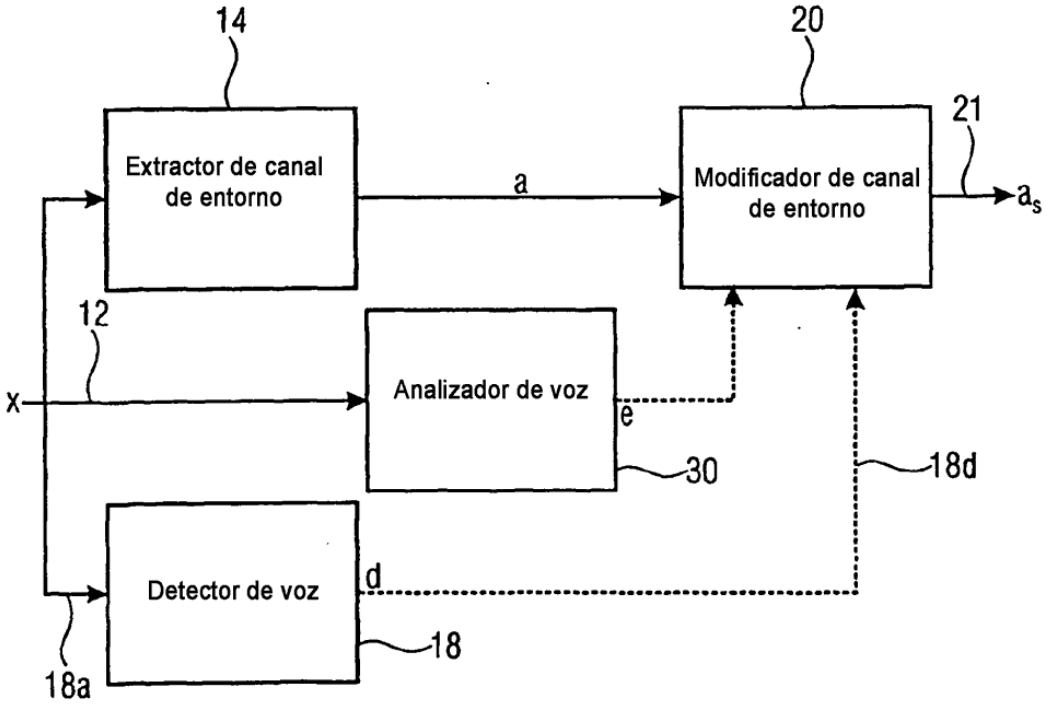

FIGURA 6D

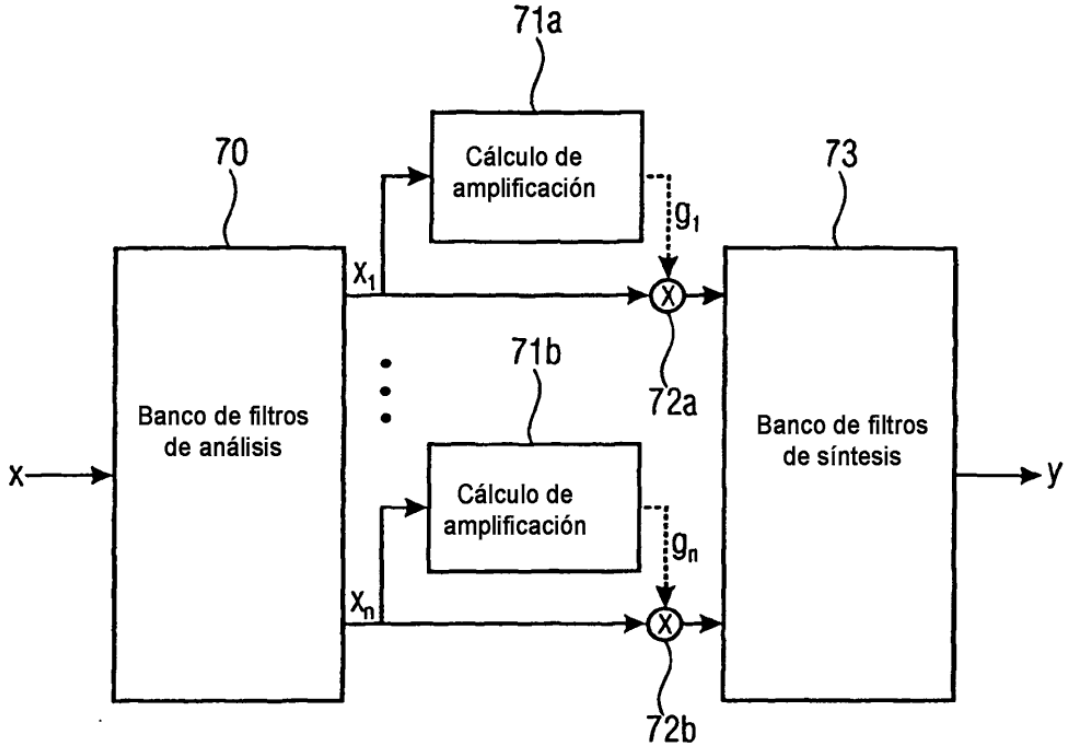

FIGURA 7

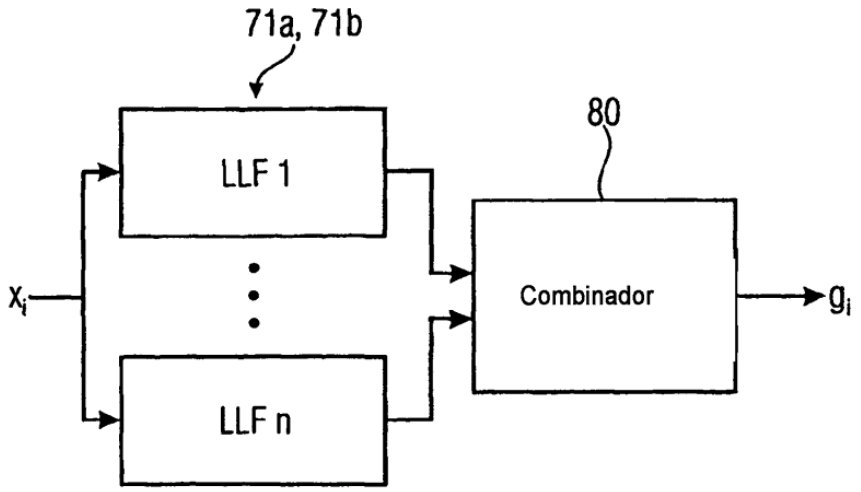

FIGURA 8