(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 391 228**

(51) Int. CI.:
G10L 11/02 (2006.01)
G10L 21/02 (2006.01)
H04R 25/00 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **08725831 .5**
(96) Fecha de presentación: **20.02.2008**
(97) Número de publicación de la solicitud: **2118885**
(97) Fecha de publicación de la solicitud: **18.11.2009**

(54) Título: **Realce de voz en audio de entretenimiento**

(30) Prioridad:
26.02.2007 US 903392 P

(45) Fecha de publicación de la mención BOPI:
22.11.2012

(45) Fecha de la publicación del folleto de la patente:
22.11.2012

(73) Titular/es:
DOLBY LABORATORIES LICENSING CORPORATION (100.0%)
100 POTRERO AVENUE
SAN FRANCISCO, CA 94103-4813, US

(72) Inventor/es:
MUESCH, HANNES

(74) Agente/Representante:
PÉREZ BARQUÍN, Eliana

DESCRIPCIÓN

Realce de voz en audio de entretenimiento

Campo técnico

La invención se refiere al procesamiento de señales de audio. Más específicamente, la invención se refiere al procesamiento de audio de entretenimiento, tal como audio de televisión, para mejorar la claridad e inteligibilidad de la voz, tal como audio de diálogo y narración. La invención se refiere a métodos, aparatos para realizar tales métodos, y a software almacenado en un medio legible por ordenador para hacer que un ordenador realice tales métodos.

Técnica antecedente

El entretenimiento audiovisual ha evolucionado a una secuencia apresurada de diálogo, narración, música y efectos. El elevado realismo que se puede alcanzar con las modernas tecnologías de audio de entretenimiento y los métodos de producción ha fomentado el uso de estilos de oratoria conversacional en la televisión que difieren sustancialmente de la presentación escenográfica claramente anunciada del pasado. Esta situación plantea un problema no sólo para la creciente población de espectadores de edad avanzada quienes, enfrentados con capacidades disminuidas de procesamiento sensorial y de lenguaje, deben esforzarse por seguir la programación sino también para las personas con audición normal, por ejemplo, cuando escuchan bajos niveles acústicos.

Lo bien que se entiende la voz depende de varios factores. Ejemplos son el cuidado de la producción de la voz (voz clara o conversacional), el ritmo de la voz, y la audibilidad de la voz. El lenguaje hablado es notablemente enérgico y puede entenderse bajo condiciones no precisamente ideales. Por ejemplo, los oyentes con problemas de audición típicamente pueden seguir una voz clara incluso cuando no pueden oír partes del discurso debido a una agudeza auditiva disminuida. Sin embargo, a medida que el ritmo de oratoria aumenta y la producción de voz se vuelve menos precisa, la escucha y la comprensión requieren mayor esfuerzo, particularmente si partes del espectro de la voz son inaudibles.

Como las audiencias televisivas no pueden hacer nada que afecte a la claridad de la voz emitida, los oyentes con problemas de audición pueden intentar compensar la audibilidad inadecuada aumentando el volumen de escucha. Aparte de resultar desagradable para las personas de audición normal que están en la misma habitación o para los vecinos, este enfoque es sólo parcialmente eficaz. Esto es así porque la mayoría de las pérdidas auditivas no son uniformes a lo largo de la frecuencia; afectan a las altas frecuencias más que a las bajas y medias frecuencias. Por ejemplo, la capacidad típica de un varón de 70 años de oír sonidos a 6 kHz es aproximadamente 50 dB peor que la de una persona joven, a frecuencias por debajo de 1 kHz la desventaja auditiva de una persona mayor es inferior a 10 dB (ISO 7029, Acústica - Distribución estadística del umbral auditivo como una función de la edad). Aumentar el volumen eleva los sonidos de baja y media frecuencia sin aumentar significativamente su contribución a la inteligibilidad porque para esas frecuencias la audibilidad ya es adecuada. Aumentar el volumen tampoco hace mucho por vencer la pérdida auditiva significativa a altas frecuencias. Una corrección más apropiada es un control de tono, como el proporcionado por un ecualizador gráfico.

Aunque es una opción mejor que simplemente aumentar el control de volumen, un control de tono aún es insuficiente para la mayoría de las pérdidas auditivas. La gran ganancia de alta frecuencia requerida para hacer que los pasajes tenues resulten audibles para el oyente con problemas de audición es probable que sea incómodamente alta durante los pasajes de nivel alto e incluso puede sobrecargar la cadena de reproducción de audio. Una solución mejor es amplificar dependiendo del nivel de la señal, proporcionando mayores ganancias a porciones de señal de nivel bajo y menores ganancias (o ninguna ganancia en absoluto) a porciones de nivel alto. Tales sistemas, conocidos como controles automáticos de ganancia (AGC) o compresores de rango dinámico (DRC) se usan en ayudas auditivas y se ha propuesto su uso para mejorar la inteligibilidad para las personas con problemas de audición en los sistemas de telecomunicación (por ejemplo, la patente de EE.UU. 5.388.185, la patente de EE.UU. 5.539.806, y la patente de EE.UU. 6.061.431).

Como la pérdida auditiva generalmente se desarrolla gradualmente, la mayoría de los oyentes con dificultades auditivas han crecido acostumbrados a sus pérdidas. Como resultado, a menudo ponen objeciones a la calidad del audio de entretenimiento cuando es procesado para compensar sus problemas de audición. Es más probable que las audiencias con problemas de audición acepten la calidad de sonido del audio compensado cuando les proporciona un beneficio tangible, como cuando aumenta la inteligibilidad del diálogo y la narración o reduce el esfuerzo mental requerido para la comprensión. Por lo tanto, es ventajoso limitar la aplicación de la compensación de pérdida auditiva a aquellas partes del programa de audio que están dominadas por voz. Hacerlo así optimiza el compromiso entre las modificaciones de calidad de sonido potencialmente desagradables de la música y los sonidos ambiente por una parte, y los beneficios de inteligibilidad deseables por otra.

El documento US 6198830 describe un método y circuito para la amplificación de señales de entrada de una ayuda auditiva, en el que una compresión de las señales captadas por la ayuda auditiva sucede en un circuito AGC

dependiente del nivel de señal adquirible. Para asegurar una compresión de dinámica, el método y circuito implementan un análisis de señal para el reconocimiento de la situación acústica además de la adquisición del nivel de señal de la señal de entrada, y el comportamiento de la compresión de dinámica se varía de manera adaptativa basándose en el resultado del análisis de la señal.

Exposición de la invención

Según un aspecto de la invención tal como se define en las reivindicaciones independientes, la voz en el audio de entretenimiento puede realzarse procesando, en respuesta a uno o más controles, el audio de entretenimiento para mejorar la claridad e inteligibilidad de porciones de voz del audio de entretenimiento, y generando un control para el procesamiento, incluyendo la generación la caracterización de segmentos de tiempo del audio de entretenimiento como (a) voz o sin voz o (b) como probabilidad de ser voz o sin voz, y la respuesta a los cambios en el nivel del audio de entretenimiento para proporcionar un control para el procesamiento, en el que a tales cambios se les responde dentro de un periodo de tiempo más corto que los segmentos de tiempo, y un criterio de decisión de la respuesta es controlado por la caracterización. El procesamiento y la respuesta pueden operar cada uno en múltiples bandas de frecuencia correspondientes, proporcionando la respuesta un control para el procesamiento para cada una de las múltiples bandas de frecuencia.

Algunos aspectos de la invención pueden operar de una manera “anticipada” de manera que cuando hay acceso a una evolución de tiempo del audio de entretenimiento antes y después de un punto de procesamiento, y en la que la generación de un control responde a al menos algún audio después del punto de procesamiento.

Algunos aspectos de la invención pueden emplear separación temporal y / o espacial de manera que alguno del procesamiento, la caracterización o la respuesta se realicen en momentos diferentes o en lugares diferentes. Por ejemplo, la caracterización puede realizarse en un primer momento o lugar, el procesamiento y la respuesta pueden realizarse en un segundo momento o lugar, y la información acerca de la caracterización de los segmentos de tiempo puede almacenarse o transmitirse para controlar los criterios de decisión de la respuesta.

Algunos aspectos de la invención también pueden incluir la codificación del audio de entretenimiento de acuerdo con un esquema de codificación perceptiva o un esquema de codificación sin pérdidas, y la descodificación del audio de entretenimiento de acuerdo con el mismo esquema de codificación empleado por la codificación, en la que alguno del procesamiento, la caracterización, y la respuesta se realizan junto con la codificación o la descodificación. La caracterización puede realizarse junto con la codificación y el procesamiento y / o la respuesta pueden realizarse junto con la descodificación.

Según los aspectos de la invención anteriormente mencionados, el procesamiento puede operar de acuerdo con uno o más parámetros de procesamiento. El ajuste de uno o más parámetros puede ser sensible al audio de entretenimiento de manera que una métrica de inteligibilidad de la voz del audio procesado se maximice o impulse por encima de un nivel umbral deseado. Según aspectos de la invención, el audio de entretenimiento puede comprender múltiples canales de audio en los que un canal es fundamentalmente voz y el otro canal o los demás canales son fundamentalmente sin voz, en los que la métrica de inteligibilidad de la voz está basada en el nivel del canal de voz y el nivel en el otro canal o los demás canales. La métrica de inteligibilidad de la voz también puede estar basada en el nivel de ruido en un ambiente de escucha en el que se reproduce el audio procesado. El ajuste de uno o más parámetros puede ser sensible a uno o más descriptores a largo plazo del audio de entretenimiento. Ejemplos de descriptores a largo plazo incluyen el nivel medio de diálogo del audio de entretenimiento y una estimación del procesamiento ya aplicado al audio de entretenimiento. El ajuste de uno o más parámetros puede ser de acuerdo con una fórmula prescriptiva, en el que la fórmula prescriptiva relaciona la agudeza auditiva de un oyente o grupo de oyentes con el uno o más parámetros. Alternativamente, o además, el ajuste de uno o más parámetros puede ser de acuerdo con las preferencias de uno o más oyentes.

Según los aspectos de la invención anteriormente mencionados, el procesamiento puede incluir múltiples funciones actuando en paralelo. Cada una de las múltiples funciones puede operar en una de múltiples bandas de frecuencia. Cada una de las múltiples funciones puede proporcionar, individual o colectivamente, control de rango dinámico, ecualización dinámica, agudización espectral, transposición de frecuencia, extracción de voz, reducción de ruido, u otra acción de realce de voz. Por ejemplo, el control de rango dinámico puede proporcionarse mediante múltiples funciones de compresión / expansión, en las que cada una procesa una zona de frecuencia de la señal de audio.

Aparte de si el procesamiento incluye o no múltiples funciones actuando en paralelo, el procesamiento puede proporcionar control de rango dinámico, ecualización dinámica, agudización espectral, transposición de frecuencia, extracción de voz, reducción de ruido, u otra acción de realce de voz. Por ejemplo, el control de rango dinámico puede proporcionarse mediante una función o dispositivo de compresión / expansión de rango dinámico.

Un aspecto de la invención es el control del realce de voz adecuado para la compensación de la pérdida auditiva de manera que, idealmente, opere sólo sobre las porciones de voz de un programa de audio y no opere sobre las restantes porciones del programa (sin voz), no tendiendo así a cambiar el timbre (distribución espectral) o la sonoridad percibida de las restantes porciones del programa (sin voz).

Según otro aspecto de la invención, el realce de voz en audio de entretenimiento comprende analizar el audio de entretenimiento para clasificar los segmentos de tiempo del audio como si fueran voz u otro audio, y aplicar compresión de rango dinámico a una o múltiples bandas de frecuencia del audio de entretenimiento durante los segmentos de tiempo clasificados como voz.

Descripción de los dibujos

La figura 1a es un diagrama esquemático de bloques funcionales que ilustra una implementación de ejemplo de aspectos de la invención.

La figura 1b es un diagrama esquemático de bloques funcionales que muestra una implementación de ejemplo de una versión modificada de la figura 1a en la que los dispositivos y / o funciones pueden estar separados temporal y / o espacialmente.

La figura 2 es un diagrama esquemático de bloques funcionales que muestra una implementación de ejemplo de una versión modificada de la figura 1a en la que el control de realce de voz se obtiene de una manera “anticipada”.

Las figuras 3a - c son ejemplos de transformaciones de potencia a ganancia útiles para entender el ejemplo de la figura 4.

La figura 4 es un diagrama esquemático de bloques funcionales que muestra cómo la ganancia de realce de voz en una banda de frecuencia puede obtenerse a partir de la estimación de potencia de la señal de esa banda de acuerdo con aspectos de la invención.

Mejor modo de llevar a cabo la invención

Las técnicas para clasificar el audio en voz y sin voz (como la música) son conocidas en la técnica y a veces son conocidas como discriminador de voz frente a otros (“SVO”). Véanse, por ejemplo, las patentes de EE.UU. 6.785.645 y 6.570.991 así como la solicitud de patente de EE.UU. publicada 20040044525, y las referencias contenidas en las mismas. Los discriminadores de audio de voz frente a otros analizan segmentos de tiempo de una señal de audio y extraen uno o más descriptores de señal (rasgos) de cada segmento de tiempo. Tales rasgos se pasan a un procesador que produce una estimación de probabilidad de que el segmento de tiempo sea voz o toma una decisión firme sobre voz / sin voz. La mayoría de los rasgos reflejan la evolución de una señal a lo largo del tiempo. Ejemplos típicos de rasgos son el ritmo al que el espectro de la señal cambia a lo largo del tiempo o el sesgo de la distribución del ritmo al que cambia la polaridad de la señal. Para reflejar fiablemente las distintas características de la voz, los segmentos de tiempo deben ser de suficiente duración. Como muchos rasgos están basados en características de la señal que reflejan las transiciones entre sílabas adyacentes, los segmentos de tiempo cubren típicamente al menos la duración de dos sílabas (es decir, aproximadamente 250 ms) para captar una de tales transiciones. Sin embargo, los segmentos de tiempo a menudo son más largos (por ejemplo, por un factor de aproximadamente 10) para conseguir estimaciones más fiables. Aunque de funcionamiento relativamente lento, los SVO son razonablemente fiables y exactos al clasificar el audio en voz y sin voz. Sin embargo, para realzar la voz selectivamente en un programa de audio de acuerdo con aspectos de la presente invención, es deseable controlar el realce de voz a una escala de tiempo más precisa que la duración de los segmentos de tiempo analizados por un discriminador de voz frente a otros.

Otra clase de técnicas, a veces conocidas como detectores de actividad vocal (VAD), indican la presencia o ausencia de voz en un fondo de ruido relativamente uniforme. Los VAD se usan extensamente como parte de esquemas de reducción de ruido en aplicaciones de comunicación por voz. A diferencia de los discriminadores de voz frente a otros, los VAD normalmente tienen una resolución temporal que es adecuada para el control de realce de voz de acuerdo con aspectos de la presente invención. Los VAD interpretan un aumento súbito de la potencia de la señal como el principio de un sonido de voz y una disminución súbita de la potencia de la señal como el final de un sonido de voz. Al hacerlo así, señalan la demarcación entre voz y fondo casi instantáneamente (es decir, dentro de una ventana de integración temporal para medir la potencia de la señal, por ejemplo, aproximadamente 10 ms). Sin embargo, como los VAD reaccionan a cualquier cambio súbito de la potencia de la señal, no pueden diferenciar entre voz y otras señales dominantes, como música. Por lo tanto, si se usan solos, los VAD no son adecuados para controlar el realce de voz para realzar la voz selectivamente de acuerdo con la presente invención.

Un aspecto de la invención es combinar la especificidad de la voz frente a la sin voz de los discriminadores de voz frente a otros (SVO) con la agudeza temporal de los detectores de actividad vocal (VAD) para facilitar el realce de voz que responda selectivamente a la voz en una señal de audio con una resolución temporal que sea más precisa que la encontrada en los discriminadores de voz frente a otros de la técnica anterior.

Aunque, en principio, los aspectos de la invención pueden implementarse en los dominios analógico y / o digital, es probable que las implementaciones prácticas se implementen en el dominio digital en el que cada una de las señales de audio está representada por muestras individuales o muestras dentro de bloques de datos.

Haciendo referencia ahora a la figura 1a, se muestra un diagrama esquemático de bloques funcionales que ilustra aspectos de la invención en el que una señal de entrada de audio 101 se pasa a una función o dispositivo de realce de voz ("Realce de voz") 102 que, cuando se lo permite una señal de control 103, produce una señal de salida de audio con voz realzada 104. La señal de control es generada por una función o dispositivo de control ("Controlador de realce de voz") 105 que opera sobre segmentos de tiempo almacenados en memoria intermedia de la señal de entrada de audio 101. El controlador de realce de voz 105 incluye una función o dispositivo discriminador de voz frente a otros ("SVO") 107 y un conjunto de una o más funciones o dispositivos detectores de actividad vocal ("VAD") 108. El SVO 107 analiza la señal a lo largo de un intervalo de tiempo que es más largo que el analizado por el VAD. El hecho de que el SVO 107 y el VAD 108 operen a lo largo de intervalos de tiempo de diferentes duraciones se ilustra gráficamente por un corchete que accede a una zona ancha (asociada con el SVO 107) y otro corchete que accede a una zona más estrecha (asociada con el VAD 108) de una función o dispositivo de memoria intermedia de señales ("Memoria intermedia") 106. La zona ancha y la zona más estrecha son esquemáticas y no están a escala. En el caso de una implementación digital en la que los datos de audio son transportados en bloques, cada porción de la memoria intermedia 106 puede almacenar un bloque de datos de audio. La zona a la que accede el VAD incluye las porciones más recientes de la señal almacenada en la memoria intermedia 106. La probabilidad de que la sección de la señal actual sea voz, tal como se determina mediante el SVO 107, sirve para controlar 109 el VAD 108. Por ejemplo, puede controlar un criterio de decisión del VAD 108, influyendo así en las decisiones del VAD.

La memoria intermedia 106 simboliza la memoria inherente al procesamiento y puede implementarse o no directamente. Por ejemplo, si el procesamiento se realiza sobre una señal de audio que está almacenada en un medio con acceso aleatorio a la memoria, ese medio puede servir como memoria intermedia. De manera similar, la historia de la entrada de audio puede reflejarse en el estado interno del discriminador de voz frente a otros 107 y el estado interno del detector de actividad vocal, en cuyo caso no es necesaria una memoria intermedia separada.

El realce de voz 102 puede estar compuesto de múltiples dispositivos o funciones de procesamiento de audio que trabajan en paralelo para realzar la voz. Cada dispositivo o función puede operar en una zona de frecuencia de la señal de audio en la que ha de realzarse la voz. Por ejemplo, los dispositivos o funciones pueden proporcionar, individualmente o en conjunto, control de rango dinámico, ecualización dinámica, agudización espectral, transposición de frecuencia, extracción de voz, reducción de ruido u otra acción de realce de voz. En los ejemplos detallados de aspectos de la invención, el control de rango dinámico proporciona compresión y / o expansión en las bandas de frecuencia de la señal de audio. Así, por ejemplo, el realce de voz 102 puede ser un banco de compresores / expansores o funciones de compresión / expansión de rango dinámico, en las que cada una procesa una zona de frecuencia de la señal de audio (un compresor / expansor o una función de compresión / expansión multibanda). La especificidad de frecuencia ofrecida por la compresión / expansión multibanda es útil no sólo porque permite adaptar el patrón de realce de voz al patrón de una pérdida auditiva dada, sino también porque permite responder al hecho de que en cualquier momento dado puede estar presente voz en una zona de frecuencia pero ausente en otra.

Para aprovechar totalmente la especificidad de frecuencia ofrecida por la compresión multibanda, cada banda de compresión / expansión puede controlarse mediante su propio detector o su propia función de detección de actividad vocal. En tal caso, cada detector o función de detección de actividad vocal puede señalar la actividad vocal en la zona de frecuencia asociada con la banda de compresión / expansión que controla. Aunque existen ventajas en que el realce de voz 102 esté compuesto de varios dispositivos o funciones de procesamiento de audio que trabajen en paralelo, realizaciones sencillas de aspectos de la invención pueden emplear un realce de voz 102 que esté compuesto solamente de un único dispositivo o función de procesamiento de audio.

Aun cuando hay muchos detectores de actividad vocal, puede haber solamente un discriminador de voz frente a otros 107 que genere una única salida 109 para controlar todos los detectores de actividad vocal que estén presentes. La elección de usar solamente un discriminador de voz frente a otros refleja dos observaciones. Una es que el ritmo al que el patrón de actividad vocal a través de la banda cambia con el tiempo es típicamente más rápido que la resolución temporal del discriminador de voz frente a otros. La otra observación es que los rasgos usados por el discriminador de voz frente a otros se obtienen típicamente de las características espectrales que mejor pueden observarse en una señal de banda ancha. Ambas observaciones hacen que resulte poco práctico el uso de discriminadores de voz frente a otros específicos de bandas.

Una combinación de SVO 107 y VAD 108 tal como se ilustra en el controlador de realce de voz 105 también puede usarse a efectos distintos de realzar la voz, por ejemplo para estimar la sonoridad de la voz en un programa de audio, o para medir el ritmo de oratoria.

El esquema de realce de voz recién descrito puede emplearse de muchas maneras. Por ejemplo, todo el esquema puede implementarse dentro de una televisión o de un receptor digital multimedia para operar sobre la señal de audio recibida de una emisión de televisión. Alternativamente, puede estar integrado con un codificador de audio perceptivo (por ejemplo, AC-3 o AAC) o puede estar integrado con un codificador de audio sin pérdidas.

El realce de voz de acuerdo con aspectos de la presente invención puede ejecutarse en diferentes momentos o en

diferentes lugares. Consideremos un ejemplo en el que el realce de voz está integrado o asociado con un codificador de audio o un proceso de codificación. En tal caso, la porción del discriminador de voz frente a otros (SVO) 107 del controlador de realce de voz 105, que a menudo es cara en términos de cálculo, puede estar integrada o asociada con el codificador de audio o el proceso de codificación. La salida del SVO 109, por ejemplo un indicador que indica la presencia de voz, puede estar incorporada en la corriente de audio codificada. Tal información incorporada en una corriente de audio codificada a menudo se denomina como metadatos. El realce de voz 102 y el VAD 108 del controlador de realce de voz 105 pueden estar integrados o asociados con un descodificador de audio y operar sobre el audio codificado previamente. El conjunto de uno o más detectores de actividad vocal (VAD) 108 también usa la salida 109 del discriminador de voz frente a otros (SVO) 107, que extrae de la corriente de audio codificada.

La figura 1b muestra una implementación de ejemplo de tal versión modificada de la figura 1a. Los dispositivos o funciones de la figura 1b que corresponden a los de la figura 1a llevan los mismos números de referencia. La señal de entrada de audio 101 se pasa a un codificador o función de codificación ("Codificador") 110 y a una memoria intermedia 106 que cubre el intervalo de tiempo requerido por el SVO 107. El codificador 110 puede ser parte de un sistema de codificación perceptiva o sin pérdidas. La salida del codificador 110 se pasa a un multiplexor o una función de multiplexación ("Multiplexor") 112. La salida del SVO (109 en la figura 1a) se muestra que está aplicada 109a al codificador 110 o, alternativamente, aplicada 109b al multiplexor 112 que también recibe la salida del codificador 110. La salida del SVO, como un indicador como en la figura 1a, es transportada en la salida del flujo de bits del codificador 110 (como metadatos, por ejemplo) o es multiplexada con la salida del codificador 110 para proporcionar un flujo de bits empaquetado y ensamblado 114 para su almacenamiento o transmisión a un demultiplexor o función de demultiplexación ("Demultiplexor") 116 que desempaqueta el flujo de bits 114 para pasarlo a un descodificador o función de descodificación 118. Si la salida del SVO 107 se pasó 109b al multiplexor 112, entonces se recibe 109b' del demultiplexor 116 y se pasa al VAD 108. Alternativamente, si la salida del SVO 107 se pasó 109a al codificador 110, entonces se recibe 109a' del descodificador 118. Como en el ejemplo de la figura 1a, el VAD 108 puede comprender múltiples funciones o dispositivos de actividad vocal. Una función o dispositivo de memoria intermedia de señales ("Memoria intermedia") 120 alimentado por el descodificador 118 que cubre el intervalo de tiempo requerido por el VAD 108 proporciona otra alimentación al VAD 108. La salida 103 del VAD se pasa a un realce de voz 102 que proporciona la salida de audio con voz realzada como en la figura 1a. Aunque se muestran por separado por claridad de presentación, el SVO 107 y / o la memoria intermedia 106 pueden estar integrados con el codificador 110. Igualmente, aunque se muestran por separado por claridad de presentación, el VAD 108 y / o la memoria intermedia 120 pueden estar integrados con el descodificador 118 o el realce de voz 102.

Si la señal de audio que ha de ser procesada ha sido pregrabada, por ejemplo como cuando se reproduce desde un DVD en el hogar de un consumidor o cuando se procesa fuera de línea en un entorno de emisión, el discriminador de voz frente a otros y / o el detector de actividad vocal pueden operar sobre secciones de la señal que incluyen porciones de la señal que, durante la reproducción, ocurren después de la muestra de señal o el bloque de señal actuales. Esto se ilustra en la figura 2, donde la memoria intermedia de señales simbólicas 201 contiene secciones de señal que, durante la reproducción, ocurren después de la muestra de señal o bloque de señal actuales ("anticipación"). Aunque la señal no haya sido pregrabada, aún puede usarse anticipación cuando el codificador de audio tiene un retardo de procesamiento inherente sustancial.

Los parámetros de procesamiento del realce de voz 102 pueden actualizarse en respuesta a la señal de audio procesada a una velocidad que es inferior a la velocidad de respuesta dinámica del compresor. Hay varios objetivos que se podrían perseguir al actualizarlos parámetros del procesador. Por ejemplo, el parámetro de procesamiento de la función de ganancia del procesador de realce de voz puede ajustarse en respuesta al nivel medio de voz del programa para garantizar que el cambio del espectro medio de voz a largo plazo sea independiente del nivel de voz. Para entender el efecto y la necesidad de tal ajuste, considérese el siguiente ejemplo. El realce de voz se aplica solamente a una porción de alta frecuencia de una señal. En un nivel medio de voz dado, la estimación de potencia 301 de la porción de señal de alta frecuencia da un promedio P1, donde P1 es mayor que la potencia umbral de compresión 304. La ganancia asociada con esta estimación de potencia es G1, que es la ganancia media aplicada a la porción de alta frecuencia de la señal. Como la porción de baja frecuencia no recibe ganancia, el espectro medio de voz es conformado para que sea G1 dB más alto a las altas frecuencias que a las bajas frecuencias. Considérese ahora lo que ocurre cuando el nivel medio de voz aumenta una cierta cantidad, ΔL. Un aumento del nivel medio de voz en ΔL aumenta la estimación de potencia media 301 de la porción de señal de alta frecuencia hasta $P2 = P1 + \Delta L$. Como puede observarse a partir de la figura 3a, la estimación de potencia más alta P2 da origen a una ganancia, G2, que es menor que G1. Por consiguiente, el espectro medio de voz de la señal procesada muestra menor énfasis de alta frecuencia cuando el nivel medio de la entrada es alto que cuando es bajo. Como los oyentes compensan las diferencias de nivel medio de voz con su control de volumen, no es deseable la dependencia de nivel del énfasis medio de alta frecuencia. Puede eliminarse modificando la curva de ganancia de las figuras 3a - c en respuesta al nivel medio de voz. Las figuras 3a - c se analizan más adelante.

Los parámetros de procesamiento del realce de voz 102 también pueden ajustarse para garantizar que una métrica de inteligibilidad de la voz se maximice o se impulse por encima de un nivel umbral deseado. La métrica de inteligibilidad de la voz puede calcularse a partir de los niveles relativos de la señal de audio y un sonido competidor en el ambiente de escucha (como el ruido de cabina de un avión). Cuando la señal de audio es una señal de audio

multicanal con voz por un canal y señales sin voz por los canales restantes, la métrica de inteligibilidad de la voz puede calcularse, por ejemplo, a partir de los niveles relativos de todos los canales y la distribución de energía espectral en ellos. Son bien conocidas métricas de inteligibilidad adecuadas [por ejemplo, ANSI S3.5 - 1997 "Method for Calculation of the Speech Intelligibility Index" American National Standards Institute, 1997; o Müsch and Buus, "Using statistical decision theory to predict speech intelligibility. I Model Structure", Journal of the Acoustical Society of America, (2001) 109, págs. 2896 - 2909].

Pueden implementarse aspectos de la invención mostrados en los diagramas de bloques funcionales de las figuras 1a y 1b y descritos en este documento, como en el ejemplo de las figuras 3a - c y 4. En este ejemplo, la amplificación de compresión de conformación de frecuencia de los componentes de voz y la liberación del procesamiento para los componentes sin voz puede realizarse a través de un procesador de rango dinámico multibanda (no mostrado) que implementa tanto características compresivas como expansivas. Tal procesador puede estar caracterizado por un conjunto de funciones de ganancia. Cada función de ganancia relaciona la potencia de entrada en una banda de frecuencia con una ganancia de banda correspondiente, que puede aplicarse a los componentes de la señal en esa banda. En las figuras 3a - c se ilustra una de tales relaciones.

Haciendo referencia a la figura 3a, la estimación de la potencia de entrada de la banda 301 se relaciona con una ganancia deseada de la banda 302 por una curva de ganancia. Esa curva de ganancia se toma como el mínimo de dos curvas constituyentes. Una curva constituyente, mostrada por la línea continua, tiene una característica de compresión con una relación de compresión ("CR") 303 escogida apropiadamente para las estimaciones de potencia 301 por encima de un umbral de compresión 304 y una ganancia constate para estimaciones de potencia por debajo del umbral de compresión. La otra curva constituyente, mostrada por la línea discontinua, tiene una característica expansiva con una relación de expansión ("ER") 305 escogida apropiadamente para estimaciones de potencia por encima del umbral de expansión 306 y una ganancia de cero para estimaciones de potencia por debajo. La curva de ganancia final se toma como el mínimo de estas dos curvas constituyentes.

El umbral de compresión 304, la relación de compresión 303, y la ganancia en el umbral de compresión son parámetros fijos. Su elección determina cómo se procesan la envolvente y el espectro de la señal de voz en una banda particular. Idealmente, se seleccionan según una fórmula prescriptiva que determina ganancias y relaciones de compresión apropiadas en bandas respectivas para un grupo de oyentes dada su agudeza auditiva. Un ejemplo de tal fórmula prescriptiva es la NAL-NL1, que fue desarrollada por el National Acoustics Laboratory, Australia, y es descrita por H. Dillon en el documento "Prescribing hearig aid performance" [H. Dillo (Ed.), Hearing Aids (págs. 249 - 261); Sydney; Boomerang Press, 2001]. Sin embargo, también pueden basarse simplemente en la preferencia del oyente. El umbral de compresión 304 y la relación de compresión 303 en una banda particular pueden depender además de parámetros específicos de un programa de audio dado, como el nivel medio de diálogo en una banda sonora de una película.

Mientras que el umbral de compresión puede ser fijo, el umbral de expansión 306 es adaptable y varía en respuesta a la señal de entrada. El umbral de expansión puede adoptar cualquier valor dentro del rango dinámico del sistema, incluyendo valores mayores que el umbral de compresión. Cuando la señal de entrada está dominada por la voz, una señal de control descrita más adelante mueve el umbral de expansión hacia niveles bajos de manera que el nivel de entrada sea superior al rango de estimaciones de potencia al que se aplica la expansión (véanse las figuras 3a y 3b). En esa condición, las ganancias aplicadas a la señal están dominadas por la característica compresiva del procesador. La figura 3b representa un ejemplo de función de ganancia que representa tal condición.

Las estimaciones de potencia de banda de la discusión precedente pueden obtenerse analizando las salidas de un banco de filtros o la salida de una transformación del dominio de tiempo a frecuencia, como la DFT (transformada discreta de Fourier), MDCT (transformada discreta del coseno modificada), o transformadas de ondículas. Las estimaciones de potencia también pueden sustituirse por medidas que están relacionadas con la intensidad de la señal como el valor medio absoluto de la señal, la energía de Teager, o por medidas perceptivas como la sonoridad. Además, las estimaciones de potencia de banda pueden ser suavizadas en el tiempo para controlar el ritmo al que cambia la ganancia.

Según un aspecto de la invención, el umbral de expansión se sitúa idealmente de manera que cuando la señal es voz el nivel de señal está por encima de la zona expansiva de la función de ganancia y cuando la señal es audio distinto de voz el nivel de la señal está por debajo de la zona expansiva de la función de ganancia. Tal como se explica más adelante, esto puede conseguirse rastreando el nivel del audio sin voz y situando el umbral de expansión en relación con ese nivel.

Ciertos rastreadores de nivel de la técnica anterior establecen un umbral por debajo del cual se aplica expansión descendente (o silenciamiento) como parte de un sistema de reducción de ruido que trata de discriminar entre el audio deseable y el ruido no deseable. Véanse, por ejemplo, las patentes de EE.UU. 3803357, 5263091, 5774557 y 6005953. En contraste, algunos aspectos de la presente invención requieren diferenciar entre voz por una parte y todas las señales de audio restantes, como música y efectos, por otra. El ruido rastreado en la técnica anterior está caracterizado por envolventes temporal y espectral que fluctúan mucho menos que las del audio deseable. Además, el ruido a menudo tiene formas espectrales distintivas que son conocidas a priori. Tales características

diferenciadoras son aprovechadas por los rastreadores de ruido en la técnica anterior. En contraste, aspectos de la presente invención rastrean el nivel de las señales de audio sin voz. En muchos casos, tales señales de audio sin voz presentan variaciones en su envolvente y la forma espectral que son al menos tan grandes como las de las señales de audio de voz. Por consiguiente, un rastreador de nivel empleado en la presente invención requiere analizar los rasgos de la señal adecuados para la distinción entre audio de voz y sin voz más que entre voz y ruido. La figura 4 muestra cómo puede obtenerse la ganancia de realce en una banda de frecuencia a partir de la estimación de potencia de la señal de esa banda. Haciendo referencia ahora a la figura 4, una representación de una señal de banda limitada 401 se pasa a un estimador de potencia o dispositivo de estimación ("Estimación de potencia") 402 que genera una estimación de la potencia de la señal 403 en esa banda de frecuencia. Esa estimación de potencia de la señal se pasa a una transformación o función de transformación de potencia a ganancia ("Curva de ganancia") 404, que puede ser de la forma del ejemplo ilustrado en las figuras 3a - c. La transformación o función de transformación de potencia a ganancia 404 genera una ganancia de banda 405 que puede usarse para modificar la potencia de la señal en la banda (no mostrado).

La estimación de potencia de la señal 403 también se pasa a un dispositivo o función ("Rastreador de nivel") 406 que rastrea el nivel de todos los componentes de la señal en la banda que no son voz. El rastreador de nivel 406 puede incluir un circuito o función de retención mínima con fugas ("Retención mínima") 407 con una tasa de fugas adaptable. Esta tasa de fugas se controla mediante una constante de tiempo 408 y tiende a ser baja cuando la potencia de la señal está dominada por la voz y alta cuando la potencia de la señal está dominada por un audio distinto de voz. La constante de tiempo 408 puede obtenerse a partir de información contenida en la estimación de la potencia de la señal 403 en la banda. Específicamente, la constante de tiempo puede estar relacionada monótonamente con la energía de la envolvente de la señal de banda en el intervalo de frecuencia entre 4 y 8 Hz. Ese rasgo puede extraerse mediante un filtro o función de filtrado de paso de banda sintonizado apropiadamente ("Paso de banda") 409.

La salida del paso de banda 409 puede estar relacionada con la constante de tiempo 408 por una función de transferencia ("Potencia a constante de tiempo") 410. La estimación de nivel de los componentes sin voz 411, que se genera mediante el rastreador de nivel 406, es la entrada a una transformada o función de transformada ("Potencia a umbral de expansión") 412 que relaciona la estimación del nivel de fondo con un umbral de expansión 414. La combinación del rastreador de nivel 406, la transformada 412 y la expansión descendente (caracterizada por la relación de expansión 305) corresponde al VAD 108 de las figuras 1a y 1b.

La transformada 412 puede ser una simple suma, es decir, el umbral de expansión 306 puede ser un número fijo de decibelios por encima del nivel estimado del audio sin voz 411. Alternativamente, la transformada 412 que relaciona el nivel de fondo estimado 411 con el umbral de expansión 306 depende de una estimación independiente de la probabilidad de que la señal de banda ancha sea voz 413. Así, cuando la estimación 413 indica una alta probabilidad de que la señal sea voz, se reduce el umbral de expansión 306. A la inversa, cuando la estimación 413 indica una baja probabilidad de que la señal sea voz, se aumenta el umbral de expansión 306. La estimación de probabilidad de voz 413 puede obtenerse a partir de un único rasgo de la señal o a partir de una combinación de rasgos de la señal que distinguen la voz de otras señales. Corresponde a la salida 109 del SVO 107 en las figuras 1a y 1b. Rasgos adecuados de la señal y métodos para procesarlos para obtener una estimación de la probabilidad de voz 413 resultan conocidos para los expertos en la materia. Se describen ejemplos en las patentes de EE.UU. 6.785.645 y 6.570.991, así como en la solicitud de patente de EE.UU. 20040044525, y en las referencias contenidas en este documento.

Se hace referencia a las siguientes patentes, solicitudes de patente y publicaciones.

Patente de Estados Unidos 3.803.357; Sacks, 9 de abril de 1974, Noise Filter.

Patente de Estados Unidos 5.263.091; Waller, Jr. 16 de noviembre de 1993, Intelligent automatic threshold circuit.

Patente de Estados Unidos 5.388.185; Terry y col., 7 de febrero de 1995, System for adaptive processing of telephone voice signals.

Patente de Estados Unidos 5.539.806; Allen y col., 23 de julio de 1996, Method for customer selection of telephone sound enhancement.

Patente de Estados Unidos 5.774.557; Slater, 30 de junio de 1998, Autotracking microphone squelch for aircraft intercom systems.

Patente de Estados Unidos 6.005.953; Stuhlfelner, 21 de diciembre de 1999, Circuit arrangement for improving the signal-to-noise ratio.

Patente de Estados Unidos 6.061.431; Knappe y col., 9 de mayo de 2000, Method for hearing loss compensation in telephony systems based on telephone number resolution.

Patente de Estados Unidos 6.570.991; Scheirer y col., 27 de mayo de 2003, Multi-feature speech / music discrimination system.

Patente de Estados Unidos 6.785.645; Khalil y col., 31 de agosto de 2004, Real-time speech and music classifier.

Patente de Estados Unidos 6.914.988; Irwan y col., 5 de julio de 2005, Audio reproducing device.

Solicitud de patente publicada de Estados Unidos 2004 / 0044525; Vinton, Mark Stuart y col., 4 de marzo de 2004, controlling loudness of speech in signals that contain speech and other types of audio material.

"Dynamic Range Control via Metadata", por Charles Q. Robinson y Kenneth Gundry, Convention Paper 5028, 107th Audio Engineering Society Convention, Nueva York, 24 - 27 de septiembre de 1999.

Implementación

La invención puede implementarse en hardware o software, o una combinación de ambos (por ejemplo, matrices lógicas programables). A menos que se especifique otra cosa, los algoritmos incluidos como parte de la invención no están relacionados inherentemente con ningún ordenador u otro aparato particular. En particular, pueden usarse diversas máquinas de propósito general con programas escritos de acuerdo con las técnicas de este documento, o puede ser más conveniente construir aparatos más especializados (por ejemplo, circuitos integrados) para realizar las etapas de método requeridas. Así, la invención puede implementarse en uno o más programas informáticos que se ejecutan en uno o más sistemas informáticos programables que comprenden cada uno al menos un procesador, al menos un sistema de almacenamiento de datos (incluyendo memoria volátil y no volátil y / o elementos de almacenamiento), al menos un dispositivo o puerto de entrada, y al menos un dispositivo o puerto de salida. El código de programa se aplica a los datos de entrada para realizar las funciones descritas en este documento y generar información de salida. La información de salida se aplica a uno o más dispositivos de salida, de manera conocida.

Cada uno de tales programas puede implementarse en cualquier lenguaje informático deseado (incluyendo lenguajes máquina, ensamblador, o de programación orientada a objetos, lógica o procedimental de alto nivel) para comunicarse con un sistema informático. En cualquier caso, el lenguaje puede ser un lenguaje compilado o interpretado.

Cada uno de tales programas informáticos se almacena o se descarga preferentemente en medios o dispositivos de almacenamiento (por ejemplo, memoria o medios de estado sólido, o medios magnéticos u ópticos) legibles por un ordenador programable de propósito general o especial, para configurar y operar el ordenador cuando los medios o el dispositivo de almacenamiento sean leídos por el sistema informático para realizar los procedimientos descritos en este documento. El sistema inventivo también puede considerarse para ser implementado como un medio de almacenamiento legible por ordenador, configurado con un programa informático, donde el medio de almacenamiento así configurado hace que un sistema informático opere de una manera específica y predefinida para realizar las funciones descritas en este documento.

Se han descrito varias realizaciones de la invención. No obstante, se comprenderá que pueden hacerse diversas modificaciones. Por ejemplo, algunas de las etapas descritas en este documento pueden ser independientes del orden, y de este modo pueden realizarse en un orden diferente al descrito.

REIVINDICACIONES

1. Un método para realzar la voz en audio de entretenimiento (101), que comprende procesar, en respuesta a uno o más controles (103), dicho audio de entretenimiento (101) para mejorar la claridad e inteligibilidad de porciones de voz del audio de entretenimiento (101), incluyendo dicho procesamiento:

- variar el nivel del audio de entretenimiento (101) en cada una de múltiples bandas de frecuencia de acuerdo con una característica de ganancia (302, 404) que relaciona el nivel de la señal de banda (403) con la ganancia (405), y

- generar un control (103, 414) para variar dicha característica de ganancia (302, 404) en cada banda de frecuencia, incluyendo dicha generación:

caracterizar segmentos de tiempo de dicho audio de entretenimiento (101) como (a) voz o sin voz o (b) como probabilidad de ser voz o sin voz, en donde dicha caracterización opera sobre una única banda ancha de frecuencia,

obtener, en cada una de dichas múltiples bandas de frecuencia, una estimación de la potencia de la señal (403),

rastrear, en cada una de dichas múltiples bandas de frecuencia, el nivel de las señales de audio sin voz (411) en la banda, siendo el tiempo de respuesta del rastreo sensible a dicha estimación de la potencia de la señal,

transformar el nivel rastreado de las señales de audio sin voz (411) en cada banda en un nivel umbral de expansión adaptable correspondiente (306, 414), e

influir en cada uno de dichos niveles umbrales de expansión adaptables correspondientes (306, 414) con el resultado de dicha caracterización para producir dicho control (103, 414) para cada banda.

2. Un método para realzar la voz en audio de entretenimiento (101), que comprende procesar, en respuesta a uno o más controles (103), dicho audio de entretenimiento (101) para mejorar la claridad e inteligibilidad de porciones de voz del audio de entretenimiento (101), incluyendo dicho procesamiento:

- variar el nivel del audio de entretenimiento (101) en cada una de múltiples bandas de frecuencia de acuerdo con una característica de ganancia (302, 404) que relaciona el nivel de la señal de banda (403) con la ganancia (405), y

- generar un control (103, 414) para variar dicha característica de ganancia (302, 404) en cada banda de frecuencia, incluyendo dicha generación:

recibir caracterizaciones de segmentos de tiempo de dicho audio de entretenimiento (101) como (a) voz o sin voz o (b) como probabilidad de ser voz o sin voz, en donde dichas caracterizaciones se refieren a una única banda ancha de frecuencia,

obtener, en cada una de dichas múltiples bandas de frecuencia, una estimación de la potencia de la señal (403),

rastrear, en cada una de dichas múltiples bandas de frecuencia, el nivel de las señales de audio sin voz (411) en la banda, siendo el tiempo de respuesta del rastreo sensible a dicha estimación de la potencia de la señal,

transformar el nivel rastreado de las señales de audio sin voz (411) en cada banda en un nivel umbral de expansión adaptable correspondiente (306, 414), e

influir en cada uno de dichos niveles umbrales de expansión adaptables correspondientes (306, 414) con el resultado de dicha caracterización para producir dicho control (103, 414) para cada banda.

3. Un método según la reivindicación 1 o la reivindicación 2 en el que existe acceso a una evolución temporal del audio de entretenimiento antes y después de un punto de procesamiento, y en el que dicha generación de un control responde a al menos algún audio después del punto de procesamiento.

4. Un método según una cualquiera de las reivindicaciones 1 - 3 en el que dicho procesamiento opera de acuerdo con uno o más parámetros de procesamiento.

5. Un método según la reivindicación 4 en el que el ajuste de uno o más parámetros es sensible al audio de entretenimiento de manera que una métrica de inteligibilidad de la voz del audio procesado se maximiza o se impulsa por encima de un nivel umbral deseado.

6. Un método según la reivindicación 5 en el que el audio de entretenimiento comprende múltiples canales de audio en los que un canal es fundamentalmente voz y el otro canal o los demás canales son fundamentalmente sin voz, en los que la métrica de inteligibilidad de la voz está basada en el nivel del canal de voz y el nivel en el otro canal o los demás canales.

7. Un método según la reivindicación 5 o la reivindicación 6 en el que la métrica de inteligibilidad de la voz también está basada en el nivel de ruido en un ambiente de escucha en el que se reproduce el audio procesado.

8. Un método según una cualquiera de las reivindicaciones 4 - 7 en el que el ajuste de uno o más parámetros es sensible a uno o más descriptores a largo plazo del audio de entretenimiento.

9. Un método según la reivindicación 8 en el que un descriptor a largo plazo es el nivel medio de diálogo del audio de entretenimiento.

10. Un método según la reivindicación 8 o la reivindicación 9 en el que un descriptor a largo plazo es una estimación del procesamiento ya aplicado al audio de entretenimiento.

11. Un método según la reivindicación 4 en el que el ajuste de uno o más parámetros es de acuerdo con una fórmula prescriptiva, en el que la fórmula prescriptiva relaciona la agudeza auditiva de un oyente o grupo de oyentes con el uno o más parámetros.

12. Un método según la reivindicación 4 en el que el ajuste de uno o más parámetros es de acuerdo con las preferencias de uno o más oyentes.

13. Un método según una cualquiera de las reivindicaciones 1 - 12 en el que dicho procesamiento proporciona control de rango dinámico, ecualización dinámica, agudización espectral, extracción de voz, reducción de ruido, u otra acción de realce de voz.

14. Un método según la reivindicación 13 en el que el control de rango dinámico se proporciona mediante una función de compresión / expansión de rango dinámico.

15. Aparato que comprende medios adaptados para realizar el método de una cualquiera de las reivindicaciones 1 a 14.

16. Un programa informático, almacenado en un medio legible por ordenador para hacer que un ordenador realice el método de una cualquiera de las reivindicaciones 1 a 14.

17. Un medio legible por ordenador que almacena en el mismo el programa informático que realiza el método de una cualquiera de las reivindicaciones 1 - 14.

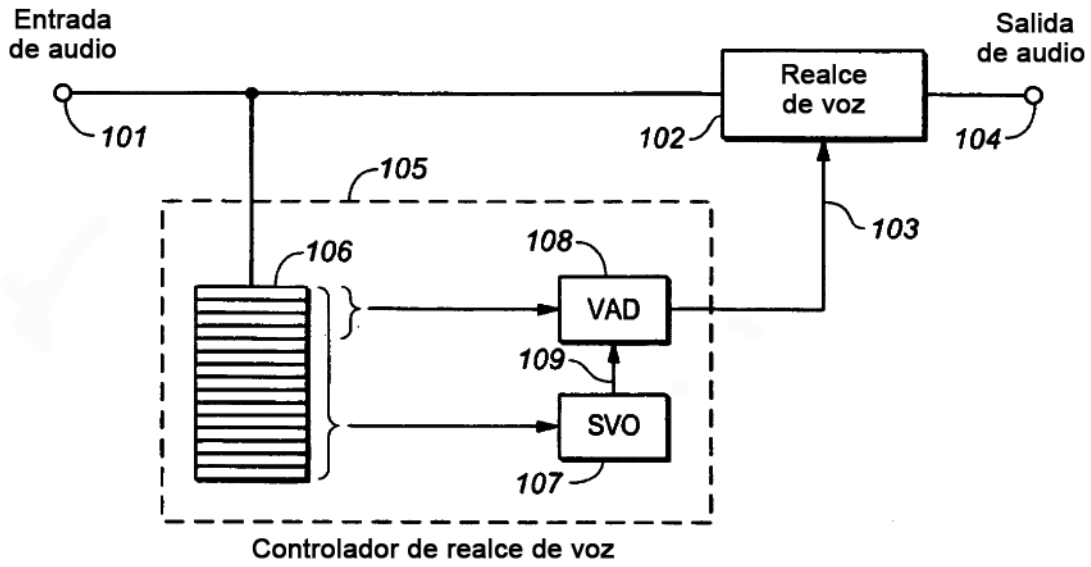
Entrada de audio
101
Realce de voz
102
Salida de audio
104
105
106
108
VAD
109
SVO
107
103
Controlador de realce de voz

FIG. 1a

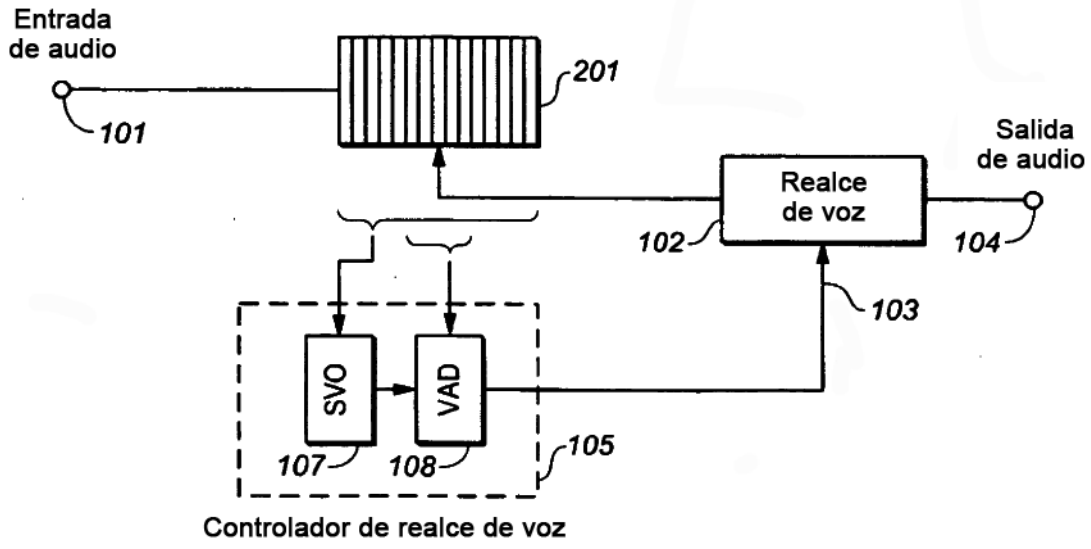
Entrada de audio
101
201
Realce de voz
102
Salida de audio
104
103
SVO
VAD
107
108
105
Controlador de realce de voz

FIG. 2

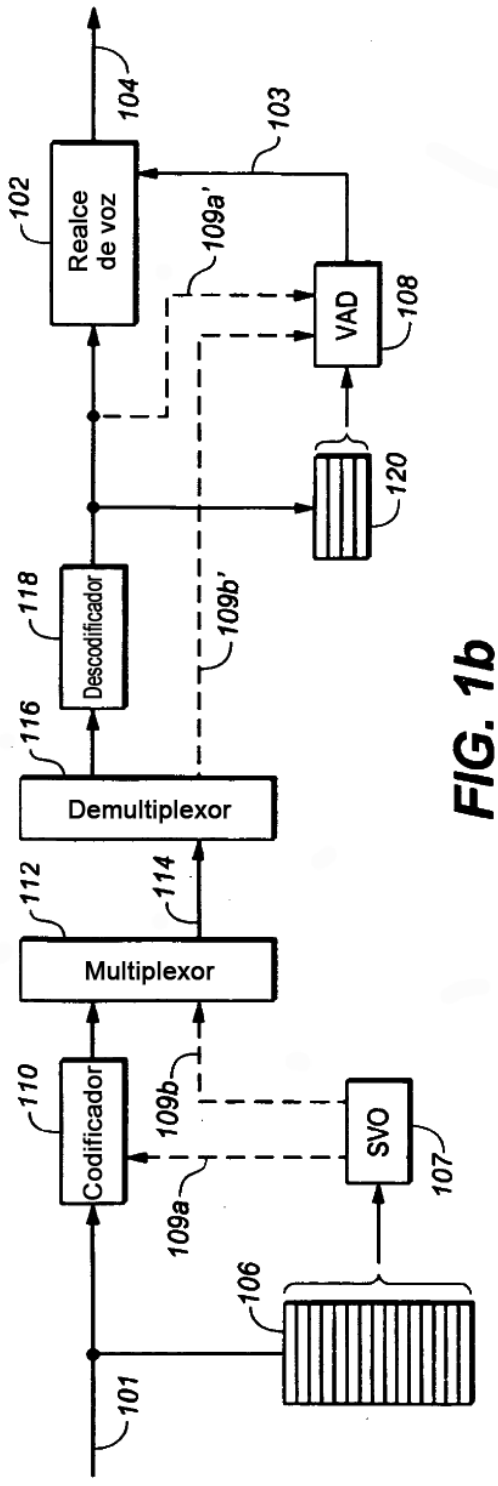
104
103
102
Realce de voz
109a'
VAD
108
120
118
Descodificador
109b'
116
Demultiplexor
112
114
Multiplexor
110
Codificador
109b
SVO
107
109a
106
101

FIG. 1b

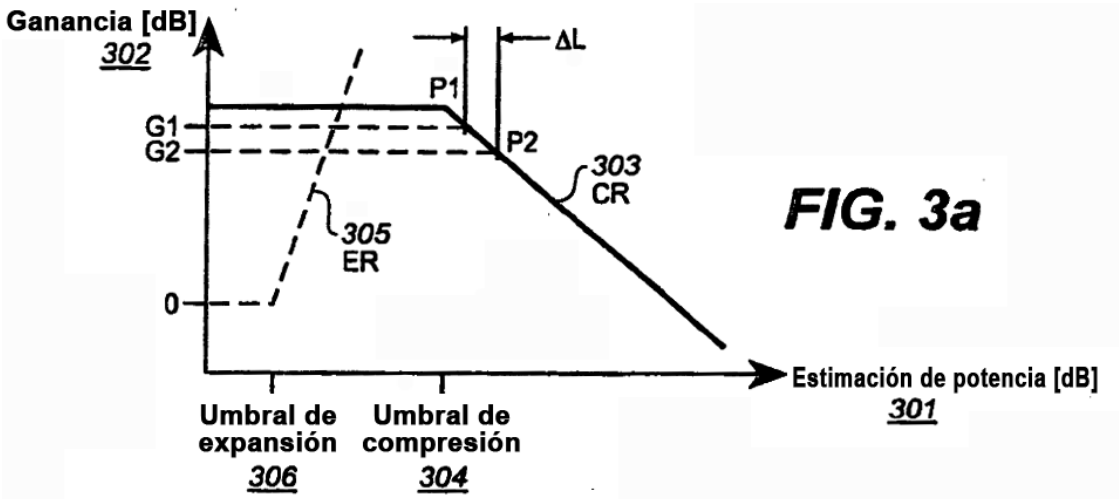

FIG. 3a

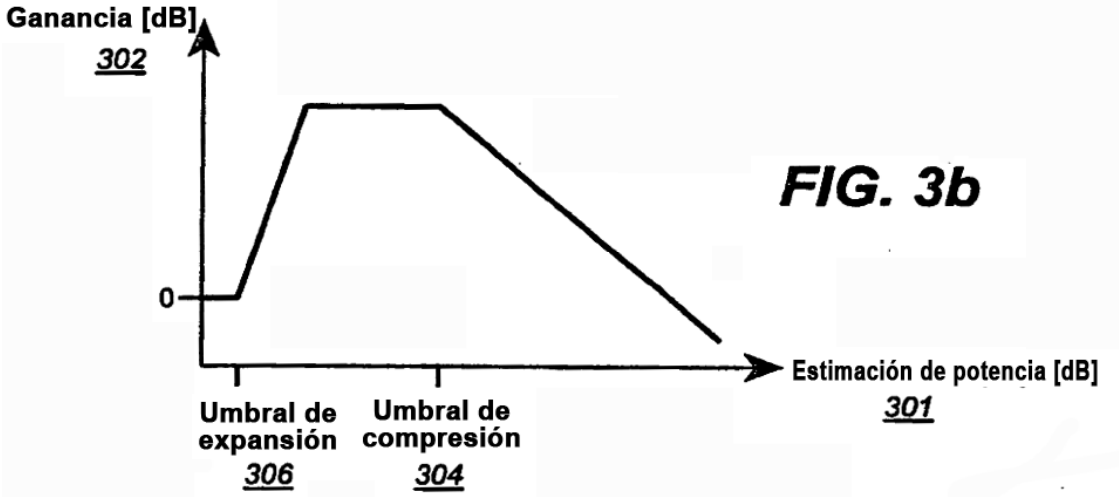

FIG. 3b

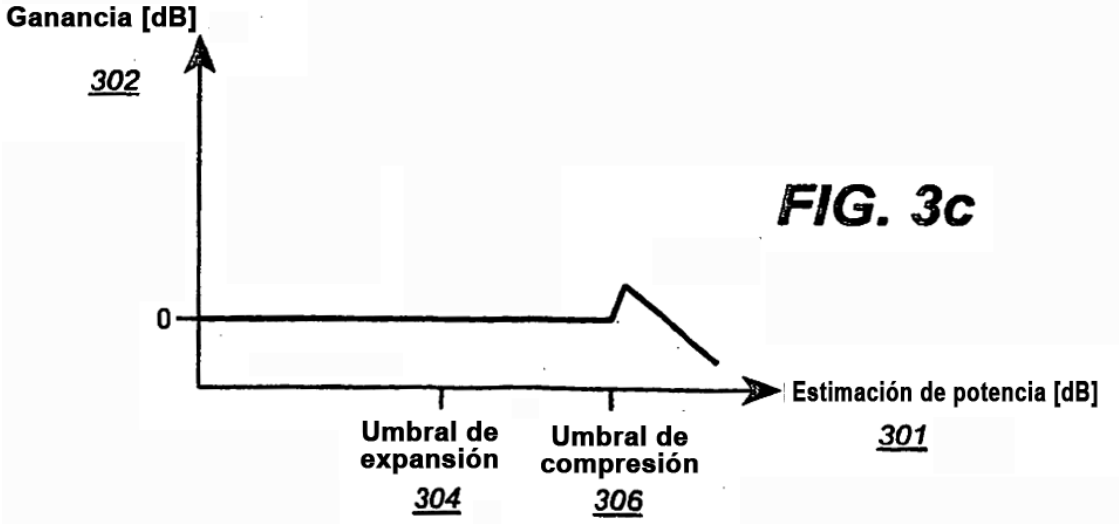

FIG. 3c

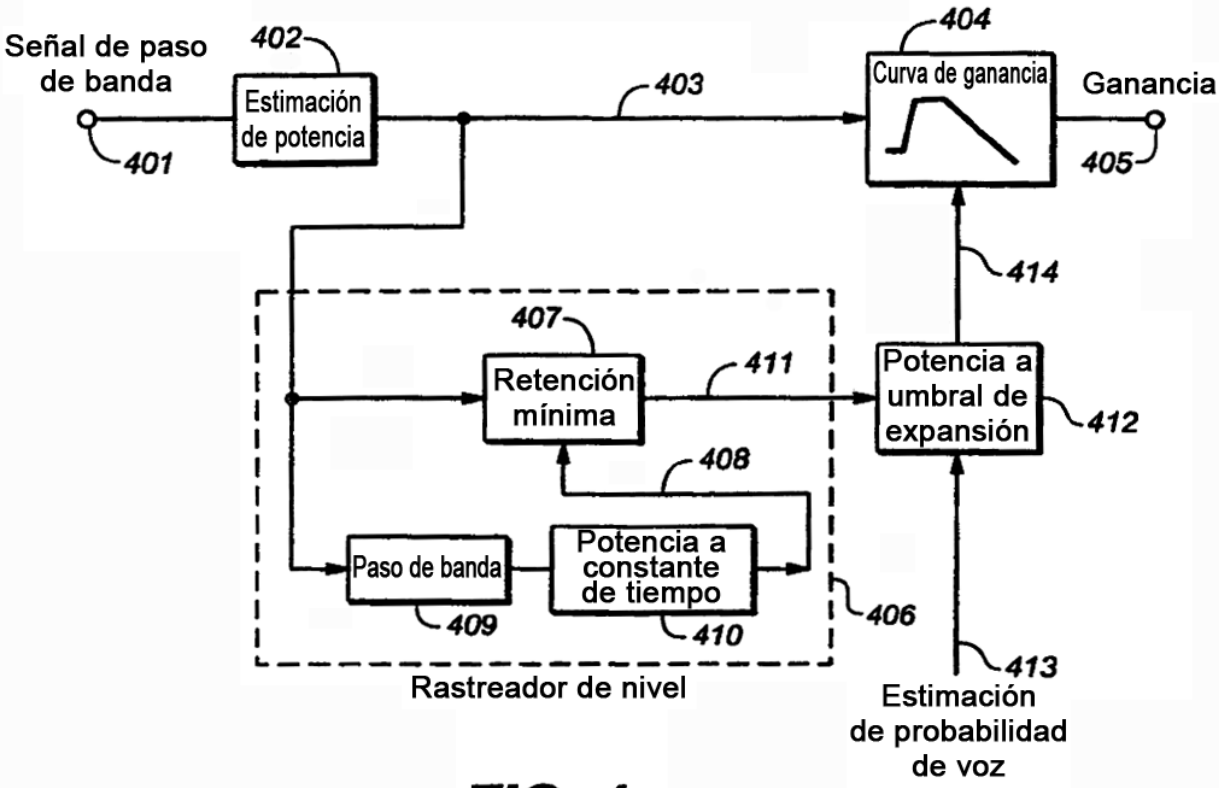
Señal de paso de banda
401
402
Estimación de potencia
403
404
Curva de ganancia
Ganancia
405
414
407
Retención mínima
411
Potencia a umbral de expansión
412
408
Paso de banda
409
Potencia a constante de tiempo
410
406
Rastreador de nivel
413
Estimación de probabilidad de voz

FIG. 4