(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 582 232**

(51) Int. CI.:

G10L 25/78 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **25.06.2009 E 09774127 (6)**

(97) Fecha y número de publicación de la concesión europea: **11.05.2016 EP 2297727**

(54) Título: **Detector de actividad de voz de múltiples micrófonos**

(30) Prioridad:

30.06.2008 US 77087 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
09.09.2016

(73) Titular/es:

DOLBY LABORATORIES LICENSING CORPORATION (100.0%)
100 Potrero Avenue
San Francisco, CA 94103-4813, US

(72) Inventor/es:

YU, RONGSHAN

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Detector de actividad de voz de múltiples micrófonos

Tecnología

La presente invención se refiere a detectores de actividad de voz. Más en particular, las formas de realización de la presente invención se refieren a detectores de actividad de voz que usan dos o más micrófonos.

Antecedentes

A no ser que se indique lo contrario en el presente documento, los enfoques descritos en esta sección no forman parte de la técnica anterior en relación con las reivindicaciones de esta solicitud ni se considera que forman parte de la técnica anterior por el hecho de que se incluyan en este apartado.

Una función de un detector de actividad de voz (VAD) es detectar la presencia o la ausencia del habla humana en las regiones de la señal de audio registrada por un micrófono. Los VAD llevan a cabo una función importante en muchos sistemas de procesamiento de voz, ya que diferentes mecanismos de procesamiento se usan en la señal de entrada dependiendo de si la misma contiene voz o no, según determine el módulo VAD. En estas aplicaciones, un funcionamiento preciso y robusto del VAD puede influir en el rendimiento global. Por ejemplo, en sistemas de comunicación de voz se usa habitualmente DTX (transmisión discontinua) para mejorar la eficacia del uso del ancho de banda. En un sistema de este tipo, el VAD se usa para determinar la presencia o la ausencia de voz en la señal de entrada, y la transmisión real de la señal de voz se interrumpe si no se detecta voz. En este contexto, clasificar erróneamente la voz como una perturbación puede dar como resultado la interrupción de la voz en la señal transmitida, afectando a su inteligibilidad. Por ejemplo, en un sistema de mejora de voz normalmente es necesario estimar el nivel de la señal perturbadora en la señal grabada. Esto se realiza normalmente con la ayuda de un VAD, donde el nivel de perturbación se estima a partir de las regiones que contienen solamente la señal perturbadora. Véase, por ejemplo, el documento de A. M. Kondoz, *Digital Speech Coding for Low Bit Rate Communication Systems*, capítulo 11 (John Wiley & Sons, 2004). En este caso, un VAD no preciso puede generar o bien una sobrestimación o una subestimación del nivel de perturbación, lo que finalmente puede dar lugar a una calidad de mejora de voz no del todo óptima.

En el pasado se han propuesto varios sistemas VAD. Véase, por ejemplo, el documento de A. M. Kondoz, *Digital Speech Coding for Low Bit Rate Communication Systems*, capítulo 10 (John Wiley & Sons, 2004). Algunos de estos sistemas utilizan los aspectos estadísticos de la diferencia entre la voz objetivo y la perturbación, y se basan en procedimientos de comparación de umbrales para diferenciar esa voz objetivo de las señales perturbadoras. Las mediciones estadísticas que se habían usado anteriormente en estos sistemas incluyen niveles de energía, medición del tiempo, tono, tasas de cruce por cero, medición de periodicidad, etc. La combinación de más de una medición estadística se usa en sistemas más sofisticados para mejorar adicionalmente la precisión de los resultados de detección. En general, los procedimientos estadísticos consiguen un buen rendimiento cuando la voz objetivo y la perturbación tienen características estadísticas muy diferentes, por ejemplo cuando la perturbación tiene un nivel que es estable y está por debajo del nivel de la voz objetivo. Sin embargo, en un entorno más adverso es muy complicado mantener un buen rendimiento, en particular cuando el nivel de la señal objetivo con respecto a la proporción del nivel de perturbación es bajo o la señal perturbadora presenta características similares a las de la voz.

En algunos diseños de sistemas de conformación de haz adaptativos y robustos también puede encontrarse un VAD en combinación con una disposición de micrófonos. Véase, por ejemplo, el documento de O. Hoshuyama, B. Begasse, A. Sugiyama y A. Hirano, "*A real time robust adaptive microphone array controlled by an SNR estimate*", actas de la conferencia internacional del IEEE sobre acústica, voz y procesamiento de señales, celebrada en 1998. Esos VAD se basan en la diferencia en los niveles de las diferentes salidas del sistema de conformación de haz de micrófono, donde la señal objetivo solo está presente en una salida y está bloqueada para otras salidas. La eficacia de un diseño de VAD de este tipo puede estar relacionada por tanto con la capacidad que tiene el sistema de conformación de haz de bloquear la señal objetivo para esas salidas, lo que puede resultar caro en los sistemas de uso cotidiano.

Otras referencias que pueden resultar útiles en relación con estos antecedentes, pero que no se considera que forman parte de la técnica anterior en lo que respecta a las formas de realización inventivas de ejemplo que se describirán en apartados subsiguientes, incluyen:

Referencia n.º 1: "*Digital Speech Coding for Low Bit Rate Communication Systems*", capítulo 10, de A. M. Kondoz (John Wiley & Sons, 2004);
Referencia n.º 2: "*Digital Speech Coding for Low Bit Rate Communication Systems*", capítulo 11, de A. M. Kondoz (John Wiley & Sons, 2004);
Referencia n.º 3: "*Optimal nearfield responses for microphone array*" de J.G Ryan y R. A. Goubran, actas del taller del IEEE, *Signal Processing to Audio Acoust.*, New Paltz, NY, EE.UU., 1997;

Referencia n.º 4: "*A real time robust adaptive microphone array controlled by an SNR estimate*" de O. Hoshuyama, B. Begasse, A. Sugiyama y A. Hirano, actas de la conferencia internacional del IEEE sobre acústica, voz y procesamiento de señales, celebrada en 1998;
Referencia n.º 5: US20030228023A1 / WO03083828A1 / CA2479758AA, "*Multichannel voice detection in adverse environments*"; y
Referencia n.º 6: US7174022, "*Small array microphone for beam-forming and noise suppression*".

El documento EP 0 386 765 A2 da a conocer una técnica para detectar un periodo de voz en una señal acústica. Se obtienen dos señales diferentes con relaciones de sonido/ruido diferentes. Según un procedimiento denominado "segundo procedimiento convencional", estas dos señales son señales de salida de dos micrófonos que están dispuestos o bien a dos distancias diferentes desde un altavoz, o bien delante y cerca de un lateral del altavoz. Se calcula la diferencia entre las respectivas potencias de corta duración de las dos señales. Se detecta un periodo de voz si esta diferencia es mayor que un umbral dado.

Según otro procedimiento, representado en la Fig. 9 del documento EP 0 386 765 A2, las dos señales se generan mediante una disposición de micrófonos unidireccionales y un micrófono omnidireccional, respectivamente. De nuevo, la detección de voz se lleva a cabo en función de la diferencia de potencias de corta duración. Según otro procedimiento representado en la Fig. 15 del documento EP 0 386 765 A2, se calculan los niveles de potencia respectivos de corta duración de las dos señales, y la detección de voz se lleva a cabo en función de la potencia de corta duración de la primera señal y de la diferencia de potencias de corta duración de ambas señales.

La patente estadounidense n.º 5.572.621 da a conocer un equipo de radio móvil que procesa muestras digitales de señales de voz que presentan componentes de ruido y componentes de voz. Una unidad de control determina y suaviza los valores de potencia de las muestras, y determina el mínimo de cada grupo sucesivo de un determinado número de valores de potencia suavizados. La unidad de control genera continuamente estimaciones de la relación de señal a ruido de las señales de voz basándose en el valor de potencia suavizado actual y en el valor de potencia suavizado sucesivo mínimo determinado más recientemente.

El documento WO 2007/091956 A2 da a conocer un detector de voz que procesa una única señal de entrada que se divide en una pluralidad de subseñales, donde cada una representa una subbanda de frecuencia. Para cada subseñal se calcula un valor de relación de señal de potencia/ruido conforme a una función no lineal. La suma de los valores de la relación de señal de potencia/ruido para las subseñales se calcula y se compara con un valor umbral dado.

La presente invención está definida por las reivindicaciones independientes. Las reivindicaciones dependientes se refieren a características opcionales de algunas formas de realización de la invención.

Breve descripción de los dibujos

La FIG. 1 es un diagrama que ilustra una configuración de micrófono genérica según una forma de realización de la presente invención.

La FIG. 2 es un diagrama que ilustra un dispositivo que incluye un detector de actividad de voz de micrófono dual de ejemplo según una forma de realización de la presente invención.

La FIG. 3 es un diagrama de bloques que ilustra un sistema de detector de actividad de voz de ejemplo según una forma de realización de la presente invención.

La FIG. 4 es un diagrama de flujo de un procedimiento de ejemplo de detección de actividad de voz según una forma de realización de la presente invención.

Descripción de formas de realización de ejemplo

En el presente documento se describen técnicas para la detección de actividad de voz. En la siguiente descripción se exponen, con fines explicativos, numerosos ejemplos y detalles específicos para proporcionar un entendimiento minucioso de la presente invención. Sin embargo, a los expertos en la técnica les resultará evidente que la presente invención, definida por las reivindicaciones, puede incluir algunas o todas las características de estos ejemplos, ya sea de manera individual o en combinación con otras características descritas posteriormente, y que puede incluir además modificaciones y equivalencias de las características y conceptos descritos en el presente documento.

A continuación se describen varios procedimientos y procesos. El que se describan siguiendo un determinado orden solo tiene como objetivo facilitar su exposición. Debe entenderse que etapas particulares pueden llevarse a cabo siguiendo otro orden o en paralelo, según se desee según las diversas implementaciones. El que una etapa particular deba preceder o seguir a otra se indicará de manera específica cuando no resulte evidente a partir del contexto.

Visión general

Las formas de realización de la presente invención mejoran los sistemas VAD. Según una forma de realización, se da a conocer un sistema VAD basado en una disposición de dos micrófonos. En tal forma de realización, la disposición de micrófonos está configurada de manera que un micrófono está colocado más cerca que el otro con respecto a la fuente de sonido objetivo. La decisión del VAD se realiza comparando los niveles de señal de las salidas de la disposición de micrófonos. Según una forma de realización, puede usarse más de dos micrófonos de manera similar.

Además, según un ejemplo, la presente invención incluye un procedimiento de detección de actividad de voz. El procedimiento incluye recibir una primera señal en un primer micrófono y una segunda señal en un segundo micrófono. El segundo micrófono está desplazado con respecto al primer micrófono. La primera señal incluye una primera componente objetivo y una primera componente perturbadora, y la segunda señal incluye una segunda componente objetivo y una segunda componente perturbadora. La primera componente objetivo difiere de la segunda componente objetivo en función de la distancia entre los micrófonos, y la primera componente perturbadora difiere de la segunda componente perturbadora en función de la distancia entre los micrófonos. El procedimiento incluye además estimar un primer nivel de señal en función de la primera señal, estimar un segundo nivel de señal en función de la segunda señal, estimar un primer nivel de ruido en función de la primera señal y estimar un segundo nivel de ruido en función de la segunda señal. El procedimiento incluye además calcular una primera relación en función del primer nivel de señal y el primer nivel de ruido, y calcular una segunda relación en función del segundo nivel de señal y el segundo nivel de ruido. El procedimiento incluye además calcular una decisión de actividad de voz actual en función de una diferencia entre la primera relación y la segunda relación.

Según un ejemplo, un sistema de detección de actividad de voz incluye un primer micrófono, un segundo micrófono, un estimador de nivel de señal, un estimador de nivel de ruido, un primer divisor, un segundo divisor y un detector de actividad de voz. El primer micrófono recibe una primera señal que incluye una primera componente objetivo y una primera componente perturbadora. El segundo micrófono está desplazado con respecto al primer micrófono. El segundo micrófono recibe una segunda señal que incluye una segunda componente objetivo y una segunda componente perturbadora. La primera componente objetivo difiere de la segunda componente objetivo y la primera componente perturbadora difiere de la segunda componente perturbadora en función de la distancia entre los micrófonos. El estimador de nivel de señal estima un primer nivel de señal basándose en la primera señal y estima un segundo nivel de señal basándose en la segunda señal. El estimador de nivel de ruido estima un primer nivel de ruido basándose en la primera señal y estima un segundo nivel de ruido basándose en la segunda señal. El primer divisor calcula una primera relación basándose en el primer nivel de señal y en el primer nivel de ruido. El segundo divisor calcula una segunda relación basándose en el segundo nivel de señal y en el segundo nivel de ruido. El detector de actividad de voz calcula una decisión de actividad de voz actual basándose en una diferencia entre la primera relación y la segunda relación.

Las formas de realización de la presente invención pueden llevarse a cabo como un procedimiento o un proceso. Los procedimientos pueden implementarse a través de circuitos electrónicos, como hardware, software o una combinación de los mismos. Los circuitos usados para implementar el proceso pueden ser circuitos dedicados (que llevan a cabo solamente una tarea específica) o circuitos genéricos (es decir, programados para llevar a cabo una o más tareas específicas).

Configuraciones, procesos e implementaciones de ejemplo

Según una forma de realización de la presente invención, un sistema VAD robusto determina un aspecto diferente de la diferencia entre la voz objetivo y la señal perturbadora. En muchas aplicaciones de comunicación de voz, por ejemplo teléfonos, teléfonos móviles, etc., la fuente de la voz objetivo está normalmente muy cerca del micrófono, mientras que las señales perturbadoras normalmente proceden de fuentes que están mucho más alejadas. Por ejemplo, en un teléfono móvil, la distancia entre el micrófono y la boca está en un intervalo comprendido entre 2 y 10 cm, mientras que las perturbaciones se producen normalmente a una distancia de, al menos, algunos metros del micrófono. A partir de la teoría de propagación de ondas de sonido se sabe que, en el primer caso, el nivel de la señal registrada será muy sensible a la ubicación del micrófono, de tal manera que cuanto más cerca del micrófono esté la fuente de sonido, mayor será el nivel de señal captado, y esta sensibilidad desaparece si la señal procede de un punto lejano, como en el segundo caso. Al contrario de las diferencias estadísticas descritas anteriormente, esta diferencia está relacionada con las ubicaciones geométricas de la fuente de sonido y, como resultado, es robusta y altamente predecible. Esto ofrece una característica muy robusta para diferenciar la señal de sonido objetivo de las perturbaciones.

Para aprovechar esta característica, según una forma de realización del sistema VAD se usa una disposición de dos micrófonos a pequeña escala. La disposición de micrófonos está configurada de manera que un micrófono está colocado más cerca que el otro con respecto a la fuente de sonido objetivo. Por tanto, la decisión VAD se calcula supervisando los niveles de señal de las salidas de estos dos micrófonos. La implementación detallada de una forma de realización de esta invención se describe en detalle en el resto de este documento.

Configuración de ejemplo de la disposición de micrófonos

La FIG. 1 es un diagrama de bloques que ilustra conceptualmente una configuración de una disposición de micrófonos 102 de ejemplo usada en una forma de realización de la presente invención. La disposición de micrófonos comprende dos micrófonos: un micrófono 102a (micrófono cercano) está a una distancia l_1 de la fuente de sonido objetivo 104, mientras que el otro micrófono 102b (micrófono lejano) está situado a una distancia l_2 de la fuente de sonido objetivo 104. En este caso, $l_1 < l_2$. Además, estos dos micrófonos 102a y 102b están lo bastante cerca entre sí como para considerar que están prácticamente en la misma posición desde el punto de vista de las perturbaciones lejanas. Según una forma de realización, esta condición se satisface si la distancia Δl entre estos dos micrófonos102a y 102b es de un orden u órdenes de magnitud inferior(es) en comparación con su distancia a la perturbación, lo que normalmente se cumple en aplicaciones reales en las que la disposición de los micrófonos puede tener un tamaño de varios centímetros.

Según una forma de realización, la distancia Δl entre estos dos micrófonos 102a y 102b es al menos un orden de magnitud inferior a la distancia hasta la fuente de la señal perturbadora. Por ejemplo, si se prevé que la fuente de la señal perturbadora está a un 1 metro del micrófono 102a (o 102b), la distancia Δl entre estos dos micrófonos puede ser de 2 centímetros.

Según una forma de realización, la distancia Δl entre estos dos micrófonos 102a y 102b está dentro de un orden de magnitud de la distancia hasta la fuente de la señal objetivo. Por ejemplo, si se prevé que la fuente de la señal objetivo está a 2 centímetros del micrófono 102a (o 102b), la distancia Δl entre estos dos micrófonos puede ser de 3 centímetros.

Según una forma de realización, la distancia entre el micrófono 102a (o 102b) y la fuente de la señal objetivo es un orden de magnitud mucho menor que la distancia entre el micrófono 102a (o 102b) y la fuente de la señal perturbadora. Por ejemplo, si se prevé que la fuente de la señal objetivo está a 5 centímetros del micrófono 102a (o 102b), la distancia hasta la fuente de la señal perturbadora puede ser de 51 centímetros.

En resumen, según una forma de realización, la fuente de la señal objetivo puede estar a 5 centímetros del micrófono 102a (o 102b), las perturbaciones pueden estar a al menos 1 metro del micrófono 102a (o 102b), y la distancia entre los dos micrófonos 102a y 102b puede ser de 3 centímetros.

La FIG. 2 es un diagrama de bloques que muestra un ejemplo de una disposición de micrófonos 102 que satisface los requisitos anteriores. En este caso, el micrófono cercano 102a está situado en la parte delantera de un teléfono móvil 204, y el micrófono lejano 102b está situado en la parte trasera del teléfono móvil 204. En este ejemplo particular, l_1 = 3~5 (cm), l_2 = 5~7 (cm) y Δl = 2~3 (cm).

Decisión VAD de ejemplo

La FIG. 3 es un diagrama de bloques de un sistema VAD 300 de ejemplo según una forma de realización de la presente invención. El sistema VAD 300 incluye un micrófono cercano 102a, un micrófono lejano 102b, convertidores de analógico a digital 302a y 302b, filtros de paso banda 304a y 304b, estimadores de nivel de señal 306a y 306b, estimadores de nivel de ruido 308a y 308b, divisores 310a y 310b, elementos de retardo unitario 312a y 312b, y un bloque de decisión VAD 314. Estos elementos del sistema VAD 300 llevan a cabo varias funciones, como se describe posteriormente.

En el sistema VAD 300, las salidas analógicas de la disposición de micrófonos 102 se digitalizan en señales PCM (modulación por impulsos codificados) mediante los convertidores de analógico a digital 302a y 302b. Para mejorar la robustez del algoritmo puede examinarse la gama de frecuencias que tiene una energía de voz considerable. Esto puede conseguirse procesando las señales digitalizadas con un par de filtros de paso banda (BPF) 304a y 304b, con frecuencias de paso banda que oscilan entre los 400 y los 1000 Hz.

En los bloques de estimación de nivel de señal 306a y 306b se estiman los niveles de las señales $X_i(n)$ proporcionadas por los BPF 304a y 304b. De manera conveniente, la estimación de nivel puede realizarse llevando a cabo una operación recursiva de determinación del promedio en la potencia de la señal $X_i(n)$ de la siguiente manera:

$$\sigma_i(n) = \alpha |X_i(n)|^2 + (1-\alpha)\sigma_i(n-1), \quad i = 1, 2$$

donde $0 < \alpha < 1$ es un valor bajo próximo a cero, y $\sigma_i(0)$ está inicializado a cero.

Supóngase que la señal $X_1(n)$ procede del micrófono cercano 102a y que $X_2(n)$ procede del micrófono lejano 102b. Ahora bien, si la estimación de nivel para la señal $X_1(n)$ es $\sigma_1(n) = \lambda_d(n) + \lambda_x(n)$, donde $\lambda_d(n)$ es el nivel de las componentes de la señal perturbadora y $\lambda_x(n)$ procede de la señal objetivo, el nivel de señal $X_2(n)$ viene dado por

$$\sigma_2(n) = g\left[\lambda_d(n) + p\lambda_x(n)\right]$$

En este caso, *g* es la diferencia de ganancia entre el micrófono lejano 102b y el micrófono cercano 120a; y *p* se debe al desvanecimiento de la propagación de la señal. En condiciones ideales, el nivel del sonido registrado es inversamente proporcional a la potencia de la distancia del sonido con respecto al micrófono. Véase, por ejemplo, el documento de J.G. Ryan y R. A. Goubran, "*Optimal nearfield responses for microphone array*", actas del taller del IEEE, *Signal Processing to Audio Acoust.*, (New Paltz, NY, EE.UU., 1997). En este caso, *p* viene dado por:

$$p = \left(l_1/l_2\right)^2$$

donde l_1 e l_2 son las distancias del sonido objetivo con respecto al micrófono cercano 102a y al micrófono lejano 102b, respectivamente. En aplicaciones prácticas, *p* puede depender de la configuración acústica real de la disposición de micrófonos y su valor puede obtenerse a través de mediciones. Cabe señalar que se supone que los niveles de las señales perturbadoras de los dos micrófonos son idénticos después de compensar la diferencia de ganancia de los micrófonos ya que, en este caso, la diferencia del desvanecimiento de la propagación entre estos dos micrófonos es insignificante.

El sistema VAD 300 también supervisa los niveles de la perturbación en $X_1(n)$ y $X_2(n)$ de la siguiente manera:

$$\lambda_i(n) = \begin{cases} \beta\,|X_i(n)|^2 + (1-\beta)\lambda_i(n-1) & VAD(n-1) = 0 \\ \lambda_i(n-1) & \text{en otro caso} \end{cases}, \quad i = 1, 2$$

donde $0 < \beta < 1$ es un valor bajo próximo a cero, y $\lambda_i(0)$ está inicializado a cero. En este caso, solo las muestras que se han clasificado como perturbaciones (VAD = 0) se incluyen en la estimación. Puesto que la decisión VAD de la muestra actual no ha realizado todavía, en este caso se usa la decisión VAD de la muestra anterior (a través de los retardos 312a y 312b). Asimismo, suponiendo que $\lambda_1(n) = \bar{\lambda}_d(n)$, $\lambda_2(n)$ se calculará como:

$$\lambda_2(n) = g\bar{\lambda}_d(n)$$

debido a la diferencia de ganancia entre el micrófono lejano y el micrófono cercano.

En general, $\lambda_d(n) \neq \bar{\lambda}_d(n)$, aunque ambos son niveles estimados de las perturbaciones. Esto se debe a que las constantes de tiempo usadas en estos dos estimadores de nivel (α y β) son diferentes. Normalmente, puede seleccionarse un valor mayor de α ya que es deseable que la respuesta del estimador del nivel de señal sea lo bastante rápida cuando el objetivo está presente; y un valor más pequeño de β para permitir una estimación sencilla del nivel de perturbación. Por este motivo, $\lambda_d(n)$ se denomina estimación a corto plazo del nivel de perturbación, y $\bar{\lambda}_d(n)$ se denomina estimación a largo plazo del nivel de perturbación. Según una forma de realización, α=0,1 y β=0,01. En otras formas de realización, los valores de α y β pueden ajustarse dependiendo de las características de la señal objetivo y de la señal perturbadora. Estos dos valores pueden fijarse de manera empírica, dependiendo de las características de las señales.

En el sistema VAD se calculan además las siguientes relaciones:

$$r_1(n) \triangleq \frac{\sigma_1(n)}{\lambda_1(n)} = \gamma(n) + \xi(n)$$

y

$$r_2(n) \triangleq \frac{\sigma_2(n)}{\lambda_2(n)} = \gamma(n) + p\xi(n)$$

donde $\gamma(n) \triangleq \lambda_d(n)/\bar{\lambda}_d(n)$ es la relación de la estimación a corto plazo y a largo plazo del nivel de perturbación en el micrófono cercano 102a, y $\xi(n) \triangleq \lambda_x(n)/\bar{\lambda}_d(n)$ es la relación de las estimaciones del nivel de señal objetivo y del nivel de perturbación en el micrófono cercano 102a. Debe observarse que la diferencia de ganancia *g* de micrófono no conocida se ha cancelado en estas dos relaciones.

La decisión VAD se basa realmente en la diferencia entre estas dos relaciones:

$$u(n) \triangleq r_1(n) - r_2(n)$$
$$= (1-p)\xi(n)$$

Evidentemente, las componentes de las perturbaciones distantes se han cancelado en *u(n),* dejando solamente las componentes de la señal de voz objetivo. Esto ofrecerá una indicación muy robusta de si la señal de voz objetivo está presente o no en la señal de entrada. Según una forma de realización adicional, en una implementación se determina la decisión VAD comparando el valor de $u(n)$ con un umbral preseleccionado, de la siguiente manera:

$$VAD(n) = \begin{cases} 0 & u(n) < (1-p)\xi_{\min} \\ 1 & \text{en otro caso} \end{cases}$$

donde $\xi_{\min}$ es un umbral SNR mínimo preseleccionado para la presencia de voz en el micrófono cercano 102a. El valor de $\xi_{\min}$ determina la sensibilidad del VAD, y su valor óptimo puede depender de los niveles de la voz objetivo y de la perturbación en la señal de entrada. Por lo tanto, su valor se establece mejor por medio de experimentos en las componentes específicas usadas en el VAD. Los experimentos han mostrado resultados satisfactorios fijando este umbral al valor 1.

Consideración de ejemplo para ruido generado por el viento

El ruido del viento es un tipo especial de perturbación. Puede formarse debido a la turbulencia de aire que se genera cuando el flujo de aire del viento es bloqueado por un objeto con bordes irregulares. A diferencia de algunas otras perturbaciones, el ruido del viento puede producirse en una ubicación muy próxima al micrófono, por ejemplo en los bordes del dispositivo de grabación o el micrófono. Cuando esto sucede, pueden generarse valores elevados de $u(n)$ incluso cuando la voz objetivo no está presente, dando lugar a problemas de falsa alarma. Por tanto, una forma de realización del bloque de decisión VAD 314 detecta además el ruido del viento mediante el cálculo y/o el análisis de la relación entre $r_1(n)$ y $r_2(n)$:

$$v(n) \triangleq r_1(n)/r_2(n)$$

Si el ruido del viento no está presente, esto da lugar a lo siguiente:

$$v(n) = \frac{1+\psi(n)}{1+p\psi(n)}$$

donde $\psi(n) \triangleq \lambda_x(n)/\lambda_d(n)$. Por tanto, el valor $v(n)$ adquiere un valor entre 1 y $1/p$ dependiendo del valor real de $\psi(n)$. Por otro lado, si hay ruido de viento presente, es posible que se produzca en una ubicación diferente en relación con la fuente de la voz objetivo y, por tanto, $v(n)$ puede estar fuera de su intervalo normal. Esto proporciona una indicación de la presencia del ruido del viento. En base a esto, la siguiente regla de decisión se usa en el sistema que ha demostrado ser muy robusto a la perturbación del ruido del viento:

$$VAD(n) = \begin{cases} 1 & u(n) \geq (1-p)\xi_{\min} \quad \text{Y} \quad \frac{1}{\varepsilon} < v(n) < \frac{\varepsilon}{p} \\ 0 & \text{en otro caso} \end{cases}$$

En este caso, ε es una constante ligeramente mayor que 1, que puede proporcionar un grado de tolerancia al error para el sistema VAD 300. Según una forma de realización, el valor de ε puede ser de 1,20. La selección del valor usado para ε puede ajustarse en otras formas de realización para ajustar la sensibilidad del VAD con respecto al ruido del viento.

La FIG. 4 es un diagrama de flujo de un procedimiento 400 de ejemplo según una forma de realización de la presente invención. El procedimiento 400 puede implementarse mediante, por ejemplo, el sistema de detección de actividad de voz 300 (véase la FIG. 3).

En la etapa 410, las señales de entrada al sistema son recibidas por los micrófonos. En un sistema con dos micrófonos, el primer micrófono está más cerca de la fuente de la señal objetivo (por ejemplo, la voz del usuario) que el segundo micrófono, pero la distancia hasta la fuente de la señal perturbadora (por ejemplo, el ruido) es mucho mayor que la distancia hasta la fuente de la señal objetivo más la distancia entre los micrófonos. Por ejemplo, en el sistema 300 (véase la FIG. 3), el micrófono 102a está más cerca de la fuente objetivo que el micrófono 102b, pero ambos micrófonos 102a y 102b están relativamente alejados de la fuente perturbadora (no mostrada).

En la etapa 420 se estima el nivel de la señal y el nivel de ruido en cada micrófono. Por ejemplo, en el sistema 300 (véase la FIG. 3), el estimador de nivel de señal 306a estima el nivel de señal en el primer micrófono, el estimador de nivel de ruido 308a estima el nivel de ruido en el primer micrófono, el estimador de nivel de señal 306b estima el nivel de señal en el segundo micrófono, y el estimador de nivel de ruido 308b estima el nivel de ruido en el segundo micrófono. A modo de ejemplo, un estimador de nivel combinado estima dos o más de los cuatro niveles, por ejemplo en función de una compartición de tiempo.

Como se ha descrito anteriormente con referencia a la FIG. 3, la estimación de nivel de ruido puede tener en cuenta la decisión de detección de actividad de voz anterior.

En la etapa 430 se calcula la relación del nivel de señal con respecto al nivel de ruido en cada micrófono. Por ejemplo, en el sistema 300 (véase la FIG. 3), el divisor 310a calcula la relación en el primer micrófono, y el divisor 310b calcula la relación en el segundo micrófono. A modo de ejemplo, un divisor combinado puede calcular ambas relaciones, por ejemplo según una compartición de tiempo.

En la etapa 440, la decisión de detección de actividad de voz actual se realiza según la diferencia entre las dos relaciones. Por ejemplo, en el sistema 300 (véase la FIG. 3), el detector VAD 314 indica la presencia de actividad de voz cuando la diferencia supera un umbral definido.

Cada una de las etapas descritas anteriormente puede incluir subetapas. Los detalles de las subetapas pueden ser como los descritos anteriormente con referencia a la FIG. 3 y (por brevedad) no se repiten.

Interpretación de ejemplo para la regla de decisión VAD

En principio, $u(n)$ es la diferencia entre el nivel de señal de salida entre el micrófono lejano 102b y el micrófono cercano 102a después de haberse compensado la diferencia de ganancia entre estos dos micrófonos. En efecto, esta diferencia proporciona una indicación de la energía de los eventos de sonido que se producen muy cerca del micrófono. Según una forma de realización, la diferencia se normaliza adicionalmente mediante el nivel de perturbación, de modo que solamente un sonido cercano con una energía considerable se etiquetará como la señal de voz objetivo.

El valor $r(n)$ es la relación entre el nivel de señal de salida entre el micrófono lejano 102b y el micrófono cercano 102a después de haberse compensado la diferencia de ganancia entre estos dos micrófonos. Para la señal de voz objetivo, $r(n)$ estará dentro de un intervalo normal que se determina por la configuración acústica de la disposición de micrófonos 102. Para el ruido del viento, $r(n)$ puede estar fuera de su intervalo normal. Este fenómeno se utiliza en una forma de realización del sistema VAD 300 para diferenciar el ruido del viento de la señal de voz objetivo.

Un diseño del sistema VAD 300 puede variar en cierta medida con respecto a las formas de realización de ejemplo descritas en secciones anteriores, para su implementación en varios tipos de sistemas de voz, incluyendo teléfonos móviles, auriculares con micrófono, sistemas de videoconferencia, sistemas de juegos y sistemas de protocolo de voz sobre Internet (VOIP), entre otros.

Una forma de realización de ejemplo puede incluir más de dos micrófonos. Usando la forma de realización de ejemplo mostrada en la FIG. 3 como punto de partida, la adición de otros micrófonos implica añadir una trayectoria de señal adicional (A/D, BPF, estimadores de nivel, divisor, retardo, etc.) que aplica las ecuaciones descritas anteriormente para procesar la señal para cada micrófono adicional. Siguiendo el mismo principio, la forma de realización VAD de ejemplo puede basarse en una combinación lineal de las relaciones $r_i(n)$ calculadas como antes a partir de todos los micrófonos:

$$u(n) = \sum_{i=1}^{N} a_i r_i(n)$$

donde N es el número total de micrófonos y a_i, $i = 1,...,N$ es una constante preseleccionada que cumple lo siguiente:

$$\sum_{i=1}^{N} a_i = 0$$

de modo que las componentes de las perturbaciones de campo lejano en estas relaciones se cancelan en $u(n)$.

La selección de a_i puede llevarse a cabo de manera empírica según la disposición específica de elementos en una implementación particular. Una posible selección de a_i, $i = 1,...,N$ que da como resultado un buen rendimiento es

$$a_1 = \sum_{i=2}^{N}(1-p_i), \text{ y}$$

$$a_i = p_i - 1, \; i > 1.$$

En este caso, p_i es la diferencia de nivel del sonido objetivo entre el i-ésimo micrófono y el primer micrófono debido a la propagación de la señal. Después, el bloque de decisión VAD 314 toma la decisión VAD comparando el valor de $u(n)$ con un umbral preseleccionado, como se ha descrito anteriormente.

$$VAD(n) = \begin{cases} 0 & u(n) < \left(a_1 + \sum_{i=2}^{N} a_i p_i \right) \xi_{\min} \\ 1 & \text{en otro caso} \end{cases}$$

Implementaciones de ejemplo

Las formas de realización de la presente invención pueden implementarse en hardware o en software, o en una combinación de ambos (por ejemplo, matrices lógicas programables). A menos que se indique lo contrario, los algoritmos incluidos como parte de la invención no están relacionados intrínsecamente con ningún ordenador particular ni con ningún otro aparato. En particular, pueden usarse varias máquinas de propósito general con programas escritos según las enseñanzas del presente documento, o puede ser más conveniente fabricar aparatos más especializados (por ejemplo, circuitos integrados) para llevar a cabo las etapas de procedimiento requeridas. Por tanto, la invención puede implementarse en uno o más programas informáticos que se ejecutan en uno o más sistemas informáticos programables, donde cada uno comprende al menos un procesador, al menos un sistema de almacenamiento de datos (que incluye memoria volátil y memoria 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 datos de entrada para llevar a cabo las funciones descritas en el presente documento y generar información de salida. La información de salida se aplica a uno o más dispositivos de salida, de una manera conocida.

Cada uno de estos programas puede implementarse en cualquier lenguaje informático deseado (incluyendo lenguaje máquina, lenguaje ensamblador o lenguajes procedurales de alto nivel, lógicos u orientados a objetos) para comunicarse con un sistema informático. En cualquier caso, el lenguaje puede ser un lenguaje compilado o interpretado.

Cada programa informático de este tipo se almacena preferiblemente o se descarga en un medio o dispositivo de almacenamiento (por ejemplo, una memoria o un medio de estado sólido, o un medio magnético u óptico) legible por un ordenador programable de propósito general o específico, para configurar y hacer funcionar el ordenador cuando el medio o dispositivo de almacenamiento es leído por el sistema informático para llevar a cabo los procedimientos descritos en el presente documento. También puede considerarse que el sistema inventivo puede implementarse 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 funcione de manera específica y predefinida para llevar a cabo las funciones descritas en el presente documento.

Según una forma de realización, un procedimiento para detectar la actividad de voz incluye recibir una primera señal desde un primer micrófono. La primera señal incluye una primera componente objetivo y una primera componente perturbadora. El procedimiento incluye además recibir una segunda señal desde un segundo micrófono desplazado con respecto al primer micrófono en una distancia. La segunda señal incluye una segunda componente objetivo y una segunda componente perturbadora. La primera componente objetivo difiere de la segunda componente objetivo en función de la distancia, y la primera componente perturbadora difiere de la segunda componente perturbadora en función de la distancia. El procedimiento incluye además estimar un primer nivel de señal en función de la primera señal, estimar un segundo nivel de señal en función de la segunda señal, estimar un primer nivel de ruido en función de la primera señal, y estimar un segundo nivel de ruido en función de la segunda señal. El procedimiento incluye además calcular una primera relación en función del primer nivel de señal y del primer nivel de ruido, y calcular una segunda relación en función del segundo nivel de señal y del segundo nivel de ruido. El procedimiento incluye además calcular una decisión de actividad de voz actual en función de una diferencia entre la primera relación y la segunda relación.

Según una forma de realización, el procedimiento incluye además llevar a cabo un filtrado de paso banda en la primera señal antes de estimar el primer nivel de señal, y llevar a cabo un filtrado de paso banda en la segunda señal antes de estimar el segundo nivel de señal. Una frecuencia de paso banda oscila entre los 400 y los 1000 hercios.

Según una forma de realización, la distancia entre el primer micrófono y el segundo micrófono es al menos un orden de magnitud inferior a una segunda distancia entre el primer micrófono y una fuente perturbadora de la componente

perturbadora. Según una forma de realización, la distancia entre el primer micrófono y el segundo micrófono está dentro de un orden de magnitud de una segunda distancia entre el primer micrófono y una fuente objetivo de la componente objetivo, y la distancia entre el primer micrófono y el segundo micrófono es al menos un orden de magnitud inferior a una tercera distancia entre el primer micrófono y una fuente perturbadora de la componente perturbadora. Según una forma de realización, el primer micrófono está alejado una primera distancia de una fuente objetivo de la componente objetivo y está alejado una segunda distancia de una fuente perturbadora de la componente perturbadora, y la primera distancia es un orden de magnitud mucho menor que la segunda distancia.

Según una forma de realización, estimar el primer nivel de señal incluye estimar el primer nivel de señal llevando a cabo una operación recursiva de determinación del promedio en un nivel de potencia de la primera señal.

Según una forma de realización, estimar el primer nivel de ruido incluye estimar el primer nivel de ruido llevando a cabo, como se ha indicado mediante una decisión de actividad de voz anterior, una operación recursiva de determinación del promedio en un nivel de potencia de la primera señal.

Según una forma de realización, estimar el primer nivel de señal incluye estimar el primer nivel de señal llevando a cabo una operación recursiva de determinación del promedio en un nivel de potencia de la primera señal usando una primera constante de tiempo, y estimar el primer nivel de ruido incluye estimar el primer nivel de ruido llevando a cabo, como se ha indicado mediante una decisión de actividad de voz anterior, una operación recursiva de determinación del promedio en un nivel de potencia de la primera señal usando una segunda constante de tiempo, donde la primera constante de tiempo es mayor que la segunda constante de tiempo.

Según una forma de realización, el procedimiento incluye además detectar un ruido de viento en función de una tercera relación entre la primera relación y la segunda relación, donde calcular la decisión de actividad de voz actual incluye calcular la decisión de actividad de voz actual en función del ruido del viento y de la diferencia entre la primera relación y la segunda relación.

Según una forma de realización, un procedimiento para detectar actividad de voz incluye recibir múltiples señales desde múltiples micrófonos. El procedimiento incluye además estimar múltiples niveles de señal en función de las múltiples señales (por ejemplo, se estima el nivel de señal de cada señal). El procedimiento incluye además estimar múltiples niveles de ruido en función de las múltiples señales (por ejemplo, se estima el nivel de ruido de cada señal). El procedimiento incluye además calcular múltiples relaciones en función de los múltiples niveles de señal y los múltiples niveles de ruido (por ejemplo, para una señal procedente de un micrófono particular, el nivel de señal correspondiente y el nivel de ruido correspondiente dan como resultado una relación correspondiente a ese micrófono). El procedimiento incluye además ajustar las múltiples relaciones según múltiples constantes. (A modo de ejemplo, la constante aplicada a la relación correspondiente al segundo micrófono se obtiene de la diferencia de nivel entre el primer micrófono y el segundo micrófono). El procedimiento incluye además calcular una decisión de actividad de voz actual en función de las múltiples relaciones después de haberse ajustado por las múltiples constantes.

Según una forma de realización, un aparato incluye un circuito que lleva a cabo la detección de actividad de voz. El aparato incluye un primer micrófono, un segundo micrófono, un estimador de nivel de señal, un estimador de nivel de ruido, un primer divisor, un segundo divisor y un detector de actividad de voz. El primer micrófono recibe una primera señal que incluye una primera componente objetivo y una primera componente perturbadora. El segundo micrófono está desplazado con respecto al primer micrófono en una distancia. El segundo micrófono recibe una segunda señal que incluye una segunda componente objetivo y una segunda componente perturbadora. La primera componente objetivo difiere de la segunda componente objetivo en función de la distancia, y la primera componente perturbadora difiere de la segunda componente perturbadora en función de la distancia. El estimador de nivel de señal estima un primer nivel de señal basándose en la primera señal y estima un segundo nivel de señal basándose en la segunda señal. El estimador de nivel de ruido estima un primer nivel de ruido basándose en la primera señal y estima un segundo nivel de ruido basándose en la segunda señal. El primer divisor calcula una primera relación basándose en el primer nivel de señal y en el primer nivel de ruido. El segundo divisor calcula una segunda relación basándose en el segundo nivel de señal y en el segundo nivel de ruido. El detector de actividad de voz calcula una decisión de actividad de voz actual basándose en una diferencia entre la primera relación y la segunda relación. El aparato también funciona de manera similar a lo descrito anteriormente en relación con el procedimiento.

Un medio legible por ordenador puede incluir un programa informático que controla que un procesador ejecute el procesamiento de manera similar a lo descrito anteriormente en relación con el procedimiento.

La descripción anterior ilustra varias formas de realización de la presente invención junto con ejemplos de la manera en que pueden implementarse los aspectos de la presente invención. No debe considerarse que los ejemplos y formas de realización anteriores son las únicas formas de realización, sino que se presentan para ilustrar la flexibilidad y las ventajas de la presente invención, definida por las siguientes reivindicaciones. En base a la anterior descripción y las siguientes reivindicaciones, otras disposiciones, formas de realización, implementaciones y equivalencias resultarán evidentes a los expertos en la técnica y pueden utilizarse sin apartarse del alcance de la invención, definido por las reivindicaciones.

REIVINDICACIONES

1. Un procedimiento para llevar a cabo una detección de actividad de voz, que comprende:

recibir una primera señal de un primer micrófono, incluyendo la primera señal una primera componente objetivo y una primera componente perturbadora;
recibir una segunda señal de un segundo micrófono desplazado con respecto al primer micrófono en una distancia, incluyendo la segunda señal una segunda componente objetivo y una segunda componente perturbadora, donde la primera componente objetivo difiere de la segunda componente objetivo en función de la distancia, y donde la primera componente perturbadora difiere de la segunda componente perturbadora en función de la distancia;
estimar un primer nivel de señal en función de la primera señal;
estimar un segundo nivel de señal en función de la segunda señal;
estimar un primer nivel de ruido en función de la primera señal;
estimar un segundo nivel de ruido en función de la segunda señal;
calcular una primera relación en función del primer nivel de señal y del primer nivel de ruido;
calcular una segunda relación en función del segundo nivel de señal y del segundo nivel de ruido; y
calcular una decisión de actividad de voz actual, donde la decisión de actividad de voz actual significa que no se detecta ninguna actividad de voz si una diferencia entre la primera relación y la segunda relación es más pequeña que un umbral preseleccionado, donde el umbral es (1-p) ξ_{min}, donde p es un factor de desvanecimiento de propagación y donde ξ_{min} es un umbral SNR mínimo preseleccionado para la presencia de voz en el micrófono más cercano al sonido objetivo, y donde la decisión de actividad de voz actual significa que se detecta actividad de voz si la diferencia es mayor o igual al umbral preseleccionado.

2. El procedimiento según la reivindicación 1, que comprende además:

llevar a cabo un filtrado de paso banda en la primera señal antes de estimar el primer nivel de señal; y
llevar a cabo un filtrado de paso banda en la segunda señal antes de estimar el segundo nivel de señal, donde una frecuencia de paso banda oscila entre los 400 y los 1000 hercios.

3. El procedimiento según la reivindicación 1 o la reivindicación 2, que comprende además:

detectar un ruido de viento en función de una tercera relación entre la primera relación y la segunda relación, donde calcular la decisión de actividad de voz actual comprende calcular la decisión de actividad de voz actual en función del ruido del viento y de la diferencia entre la primera relación y la segunda relación.

4. El procedimiento según una cualquiera de las reivindicaciones 1 a 3, en el que la distancia entre el primer micrófono y el segundo micrófono es al menos un orden de magnitud inferior a una segunda distancia entre el primer micrófono y una fuente perturbadora de la componente perturbadora.

5. El procedimiento según una cualquiera de las reivindicaciones 1 a 3, en el que la distancia entre el primer micrófono y el segundo micrófono está dentro de un orden de magnitud de una segunda distancia entre el primer micrófono y una fuente objetivo de la componente objetivo, y donde la distancia entre el primer micrófono y el segundo micrófono es al menos un orden de magnitud inferior a una tercera distancia entre el primer micrófono y una fuente perturbadora de la componente perturbadora.

6. El procedimiento según una cualquiera de las reivindicaciones 1 a 3, en el que el primer micrófono está alejado una primera distancia de una fuente objetivo de la componente objetivo y está alejado una segunda distancia de una fuente perturbadora de la componente perturbadora, y donde la primera distancia es un orden de magnitud mucho menor que la segunda distancia.

7. El procedimiento según una cualquiera de las reivindicaciones 1 a 6, en el que estimar el primer nivel de señal comprende estimar el primer nivel de señal llevando a cabo una operación recursiva de determinación del promedio en un nivel de potencia de la primera señal.

8. El procedimiento según una cualquiera de las reivindicaciones 1 a 7, en el que estimar el primer nivel de ruido comprende estimar el primer nivel de ruido llevando a cabo, como se ha indicado mediante una decisión de actividad de voz anterior, una operación recursiva de determinación del promedio en un nivel de potencia de la primera señal.

9. El procedimiento según una cualquiera de las reivindicaciones 1 a 6, en el que:

estimar el primer nivel de señal comprende estimar el primer nivel de señal llevando a cabo una operación recursiva de determinación del promedio en un nivel de potencia de la primera señal usando una primera constante de tiempo; y
estimar el primer nivel de ruido comprende estimar el primer nivel de ruido llevando a cabo, como se ha indicado mediante una decisión de actividad de voz anterior, una operación recursiva de determinación del

promedio en un nivel de potencia de la primera señal usando una segunda constante de tiempo, donde la primera constante de tiempo es mayor que la segunda constante de tiempo.

10. Un aparato que incluye un circuito que está configurado para llevar a cabo una detección de actividad de voz, comprendiendo el aparato:

un primer micrófono que está configurado para recibir una primera señal que incluye una primera componente objetivo y una primera componente perturbadora;
un segundo micrófono, desplazado con respecto al primer micrófono en una distancia, que está configurado para recibir una segunda señal que incluye una segunda componente objetivo y una segunda componente perturbadora, donde la primera componente objetivo difiere de la segunda componente objetivo en función de la distancia, y donde la primera componente perturbadora difiere de la segunda componente perturbadora en función de la distancia;
un estimador de nivel de señal que está configurado para estimar un primer nivel de señal basándose en la primera señal y que estima un segundo nivel de señal basándose en la segunda señal;
un estimador de nivel de ruido que está configurado para estimar un primer nivel de ruido basándose en la primera señal y que estima un segundo nivel de ruido basándose en la segunda señal;
un primer divisor que está configurado para calcular una primera relación basándose en el primer nivel de señal y en el primer nivel de ruido;
un segundo divisor que está configurado para calcular una segunda relación basándose en el segundo nivel de señal y en el segundo nivel de ruido; y
un detector de actividad de voz que está configurado para calcular una decisión de actividad de voz actual, donde la decisión de actividad de voz actual significa que no se detecta ninguna actividad de voz si una diferencia entre la primera relación y la segunda relación es más pequeña que un umbral preseleccionado, donde el umbral es (1-p) ξ_{min}, donde p es un factor de desvanecimiento de propagación y donde ξ_{min} es un umbral SNR mínimo preseleccionado para la presencia de voz en el micrófono más cercano al sonido objetivo, y donde la decisión de actividad de voz actual significa que se detecta actividad de voz si la diferencia es mayor o igual al umbral preseleccionado.

11. El aparato según la reivindicación 12, en el que el aparato está adaptado para llevar a cabo el procedimiento según una cualquiera de las reivindicaciones 2 a 9.

12. Un producto usado para llevar a cabo una detección de actividad de voz, que comprende:

un primer micrófono que está configurado para recibir una primera señal que incluye una primera componente objetivo y una primera componente perturbadora;
un segundo micrófono, desplazado con respecto al primer micrófono en una distancia, que está configurado para recibir una segunda señal que incluye una segunda componente objetivo y una segunda componente perturbadora, donde la primera componente objetivo difiere de la segunda componente objetivo en función de la distancia, y donde la primera componente perturbadora difiere de la segunda componente perturbadora en función de la distancia;
medios para estimar un primer nivel de señal basándose en la primera señal y que estiman un segundo nivel de señal basándose en la segunda señal;
medios para estimar un primer nivel de ruido basándose en la primera señal y que estiman un segundo nivel de ruido basándose en la segunda señal;
medios para calcular una primera relación en función del primer nivel de señal y del primer nivel de ruido;
medios para calcular una segunda relación en función del segundo nivel de señal y del segundo nivel de ruido;
medios para calcular una decisión de actividad de voz actual, donde la decisión de actividad de voz actual significa que no se detecta ninguna actividad de voz si una diferencia entre la primera relación y la segunda relación es más pequeña que un umbral preseleccionado, donde el umbral es (1-p) ξ_{min}, donde p es un factor de desvanecimiento de propagación y donde ξ_{min} es un umbral SNR mínimo preseleccionado para la presencia de voz en el micrófono más cercano al sonido objetivo, y donde la decisión de actividad de voz actual significa que se detecta actividad de voz si la diferencia es mayor o igual al umbral preseleccionado; y
medios para llevar a cabo el procedimiento según una cualquiera de las reivindicaciones 2 a 9.

13. Un medio tangible legible por ordenador que incluye un programa informático para llevar a cabo la detección de actividad de voz, controlando el programa informático un procesador para ejecutar el procedimiento según una cualquiera de las reivindicaciones 1 a 9.

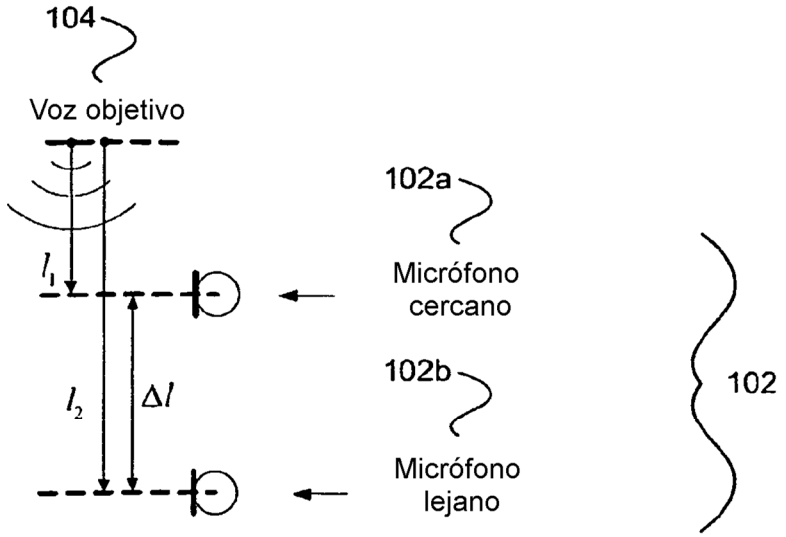
104
Voz objetivo
102a
Micrófono cercano
102b
Micrófono lejano
102
l_1
l_2
Δl

FIG. 1

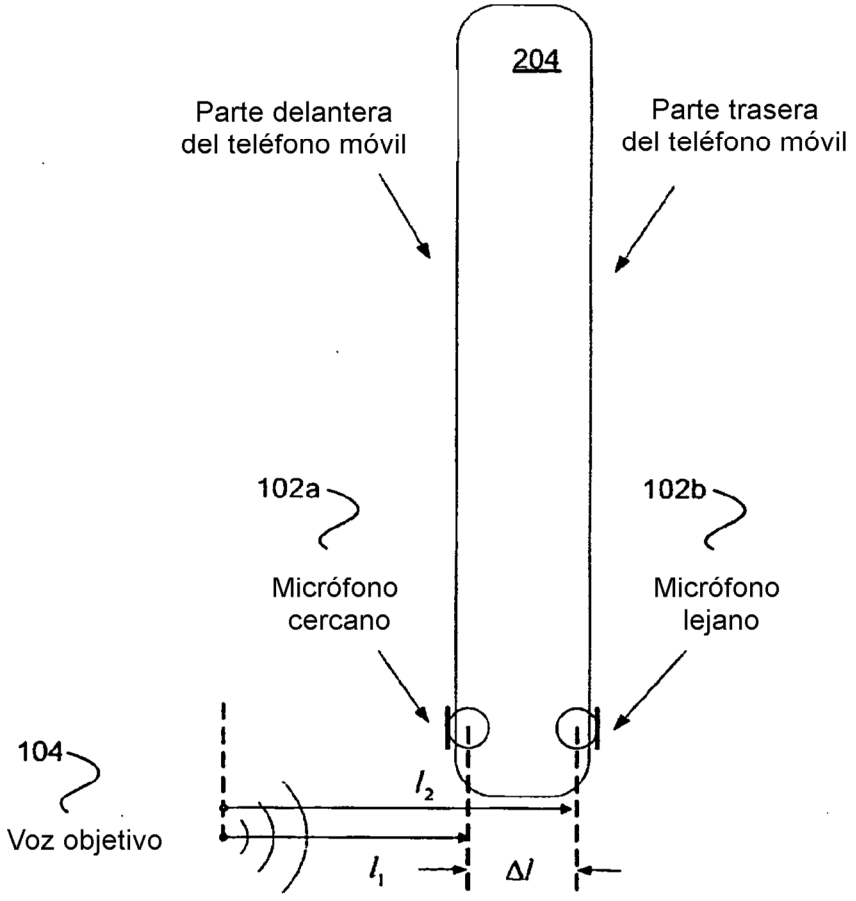
204
Parte delantera del teléfono móvil
Parte trasera del teléfono móvil
102a
Micrófono cercano
102b
Micrófono lejano
104
Voz objetivo
l_2
l_1
Δl

FIG. 2

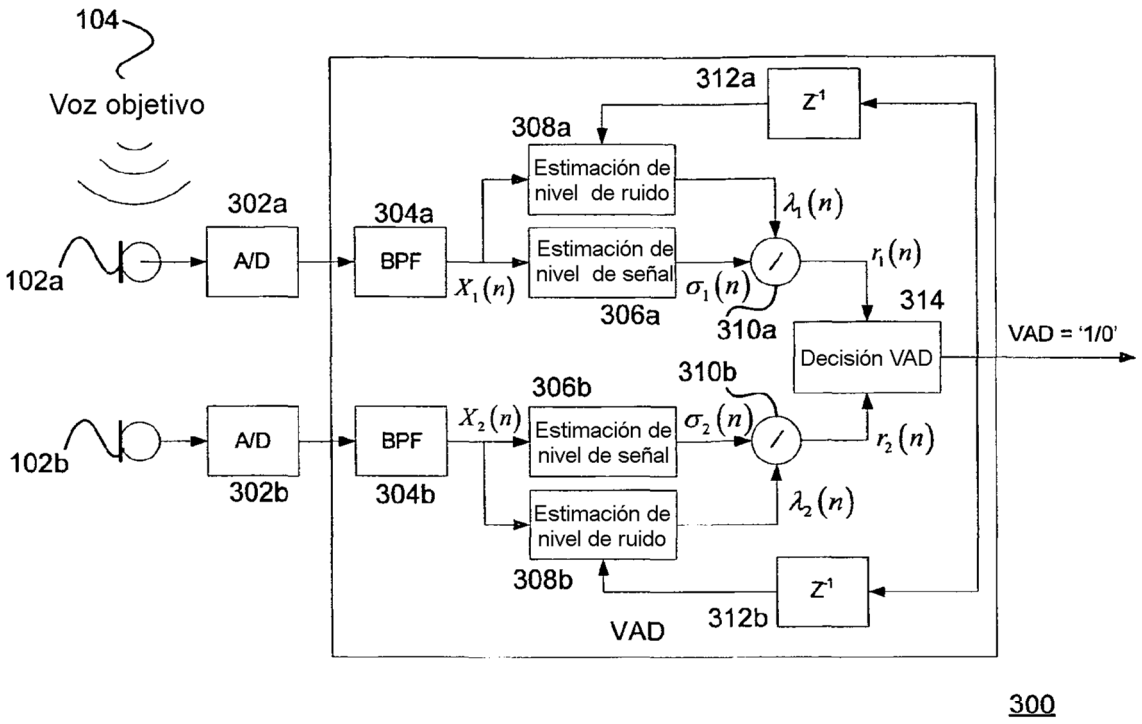
104
Voz objetivo
102a
A/D
302a
BPF
304a
Estimación de nivel de ruido
Estimación de nivel de señal
306a
$X_1(n)$
$\lambda_1(n)$
$\sigma_1(n)$
310a
$r_1(n)$
312a
z^{-1}
308a
314
Decisión VAD
VAD = '1/0'
310b
306b
$X_2(n)$
$\sigma_2(n)$
$r_2(n)$
$\lambda_2(n)$
102b
A/D
302b
BPF
304b
Estimación de nivel de señal
Estimación de nivel de ruido
308b
z^{-1}
312b
VAD
300

FIG. 3

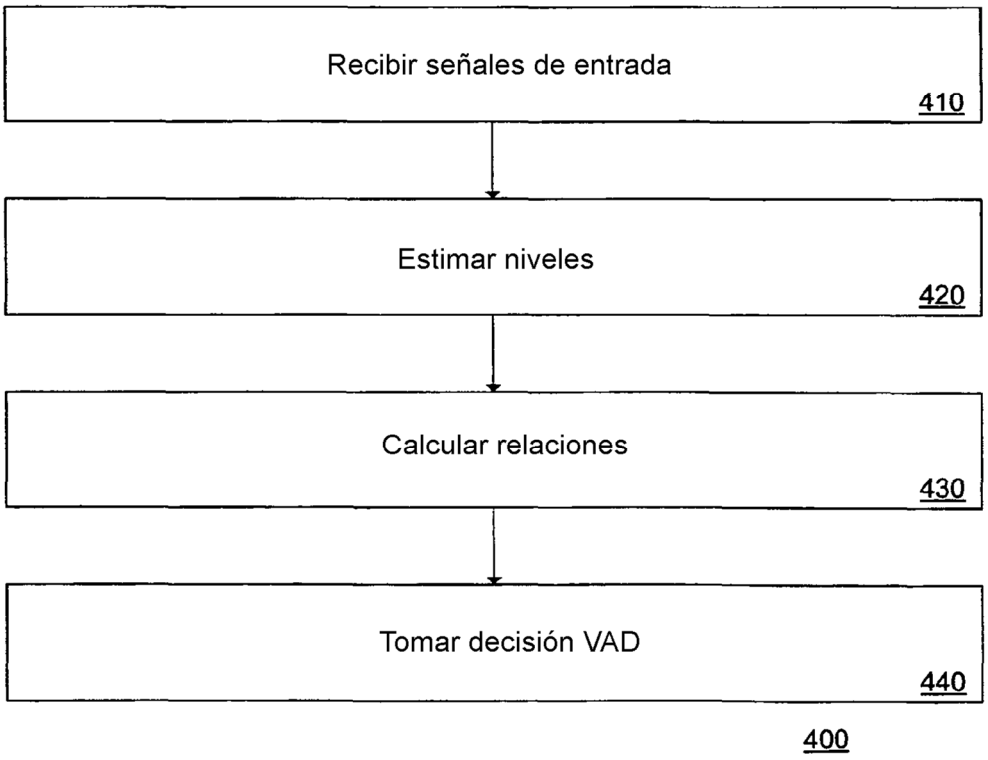
Recibir señales de entrada
410
Estimar niveles
420
Calcular relaciones
430
Tomar decisión VAD
440
400

FIG. 4