(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 746 010**

(51) Int. CI.:

G10L 25/78 (2013.01)
G10L 21/0208 (2013.01)
G10L 21/034 (2013.01)
H04R 3/00 (2006.01)
G10L 21/0216 (2013.01)
G10L 15/22 (2006.01)
G10L 15/08 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **14.09.2016 PCT/US2016/051562**

(87) Fecha y número de publicación internacional: **13.04.2017 WO17062138**

(96) Fecha de presentación y número de la solicitud europea: **14.09.2016 E 16770620 (9)**

(97) Fecha y número de publicación de la concesión europea: **17.07.2019 EP 3360137**

(54) Título: **Identificación de sonido procedente de una fuente de interés en base a múltiples suministros de audio**

(30) Prioridad:

06.10.2015 US 201514876666

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
04.03.2020

(73) Titular/es:

MICROSOFT TECHNOLOGY LICENSING, LLC (100.0%)
Attn:Patent Group Docketing, (Bldg. 8/1000), One Microsoft Way
Redmond, WA 98052-6399, US

(72) Inventor/es:

ZAD ISSA, SYAVOSH

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Identificación de sonido procedente de una fuente de interés en base a múltiples suministros de audio

Antecedentes

La identificación de sonido que se origina a partir de una fuente de interés, puede resultar problemática. Esto es especialmente cierto en presencia de ruido de fondo que puede ser de naturaleza esporádica. Los sistemas que se basan en identificación de sonido que se origina a partir de una fuente de interés, tal como, por ejemplo, un detector de actividad de voz, utilizan diversos mecanismos para intentar distinguir cuándo se origina el sonido a partir de una fuente de interés y cuándo el sonido es simplemente ruido de fondo. Estos diversos mecanismos, sin embargo, adolecen de un número de deficiencias. Una de tales deficiencias consiste en que muchos de esos diversos mecanismos son de naturaleza compleja y realizan cálculos intensivos en recursos. Como resultado, esos diversos mecanismos no son por lo general adecuados para aplicaciones de baja potencia o de bajo coste. Además, muchos de esos diversos mecanismos residen en modelos estadísticos o heurísticos que se han desarrollado mediante aprendizaje automático o mediante comparación de plantillas, lo que se añade a la complejidad de esos sistemas. El desarrollo de esos modelos estadísticos o heurísticos y los componentes correspondientes del sistema para identificar el sonido que se origina desde una fuente de interés, normalmente requiere una cantidad significativa de esfuerzo. Maj J B et al. (“Comparison of adaptive noise reduction algorithms in dual microphone hearing aids”, SPEECH COMMUNICATION, vol. 48, núm. 8, 1 de agosto de 2006) describe una evaluación física y perceptiva de dos algoritmos adaptativos de reducción de ruido para audífonos de doble micrófono. Jae-Hun Choi et al. (“Dual-microphone voice activity detection technique based on two-step power level difference ratio”, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, vol. 22, núm. 6, 1 de junio de 2014), describe una técnica de detección de actividad de voz (VAD) de doble micrófono basada en la relación de diferencia de nivel de potencia (PLD) de dos etapas.

Compendio

Este compendio se proporciona con el fin de introducir una selección de conceptos de forma simplificada, que se describen mejor más adelante en la descripción detallada. Este compendio no está destinado a identificar características clave o características esenciales del objeto reivindicado, ni está destinado a ser usado de forma aislada como ayuda a la determinación del alcance del objeto reivindicado. La presente invención está definida en las reivindicaciones independientes. Realizaciones preferidas están definidas en las reivindicaciones dependientes. Todas las menciones de la palabra “realización(es)”, si se refieren a combinaciones de características diferentes de las definidas por las reivindicaciones independientes, se refieren a ejemplos que fueron presentados originalmente pero que no representan realizaciones de la invención actualmente reivindicada; estos ejemplos se muestran solamente con fines ilustrativos.

Los ejemplos descritos en la presente memoria incluyen métodos, medios de almacenaje con ordenador, y sistemas para identificar el sonido que se origina desde una fuente de interés. En varios ejemplos, un primer suministro de audio se captura por medio de un primer micrófono de un dispositivo de computación, y un segundo suministro de audio se captura por medio de un segundo micrófono del dispositivo de computación. El primer suministro de audio puede ser procesado utilizando el segundo suministro de audio para identificar sonido que se origina a partir de la fuente de interés. Este procesamiento, en algunos ejemplos, podría incluir sincronizar en el tiempo el primer suministro de audio con el segundo suministro de audio, por ejemplo, aplicando un retardo ya sea al primer suministro de audio o ya sea al segundo suministro de audio. Este procesamiento puede incluir también atenuar, o filtrar, frecuencias del primer suministro de audio, en base a frecuencias correspondientes dentro del segundo suministro de audio. En varios ejemplos, este procesamiento puede incluir también procesar el segundo suministro de audio, utilizando el primer suministro de audio, para permitir además la identificación del sonido que se origina desde el punto de interés. De nuevo, en esas realizaciones, el procesamiento puede incluir atenuar, o filtrar, frecuencias procedentes del segundo suministro de audio, en base a frecuencias correspondientes procedentes del primer suministro de audio.

Breve descripción de los dibujos

La presente divulgación se va a describir con detalle en lo que sigue, con referencia las Figuras de los dibujos anexos:

La Figura 1 es un diagrama de bloques de un entorno operativo en el que se pueden emplear varias realizaciones de la presente descripción;

Las Figuras 2A, 2B y 2C muestran representaciones esquemáticas ilustrativas de configuraciones de sistema de procesamiento de sonido, conforme a varias realizaciones de la presente invención;

Las Figuras 3A y 3B son representaciones gráficas de niveles de confianza de fuente y niveles de confianza de ruido, conforme a varias realizaciones de la presente descripción;

La Figura 4 muestra una representación esquemática ilustrativa de un sistema de procesamiento de sonido que tiene una configuración de tres micrófonos, conforme a varias realizaciones de la presente descripción;

La Figura 5 es un diagrama de flujo que muestra un método ilustrativo para identificar sonido procedente de una fuente de interés, conforme a varias realizaciones de la presente descripción;

La Figura 6 es un diagrama de flujo que muestra un método ilustrativo para procesamiento de un primer y un segundo suministros de audio para identificar el sonido procedente de una fuente de interés, conforme a varias realizaciones de la presente descripción;

La Figura 7 es un diagrama de bloques de un entorno de computación ilustrativo, adecuado para su uso en la implementación de realizaciones descritas en la presente memoria.

Descripción detallada

El objeto de las realizaciones de la presente divulgación, se describe con especificidad en la presente memoria para cumplir los requisitos estatutarios. Sin embargo, la descripción en sí misma no se pretende que limite el alcance de esta patente. Por el contrario, los inventores han contemplado que el objeto reivindicado pueda ser también materializado de otras formas, que incluyan diferentes etapas o combinaciones de etapas similares a las que se describen en el presente documento, junto con otras tecnologías presentes o futuras. Además, aunque los términos "etapa" y/o "bloque" pueden ser usados en la presente memoria para señalar diferentes elementos de los métodos empleados, los términos no deberán ser interpretados en el sentido de que impliquen ningún orden particular entre las diversas etapas descritas en la presente memoria a menos que, y salvo cuando, el orden de las etapas individuales sea descrito de manera explícita.

A los efectos de la presente descripción, la palabra "incluyendo" tiene el mismo significado amplio que "comprendiendo", y la palabra "acceder" comprende "recibir", "hacer referencia" o "recuperar". Adicionalmente, palabras tales como "un" y "el", a menos que se indique lo contrario, incluyen tanto el plural como el singular. Por lo tanto, por ejemplo, la restricción de "una característica" se satisface donde estén presentes una o más características. También, el término "o" incluye la forma conjuntiva, la disyuntiva o ambas (*a* o *b* incluyen por tanto cualquiera de *a* o *b*, así como *a* y *b*).

A los efectos de la discusión detallada que sigue, las realizaciones se describen con referencia a un sistema para identificar sonido que se origina a partir de una fuente de lenguaje de interés; el sistema puede implementar varias componentes para llevar a cabo la funcionalidad de las realizaciones descritas en la presente memoria. Las componentes pueden estar configuradas para realizar aspectos novedosos de las realizaciones, donde "configurado para" comprende "programado para" realizar tareas particulares o implementar tipos de datos abstractos particulares que usan código. Se contempla que los métodos y los sistemas descritos en la presente memoria pueden ser llevados a cabo en diferentes tipos de entornos operativos que tengan configuraciones alternativas de los componentes funcionales. Como tal, las realizaciones descritas en la presente memoria son meramente ilustrativas, y se contempla que las técnicas puedan ser extendidas a otros contextos de implementación.

Varias realizaciones descritas en la presente memoria permiten la identificación de sonido que se origina desde una dirección de un punto de interés utilizando múltiples suministros de audio. Esto puede ser llevado a cabo procesando suministros de audio, según se describe en la presente memoria, capturados por múltiples micrófonos donde se sabe que al menos un micrófono está en relación de proximidad más cercana al punto de interés. Este procesamiento puede ayudar a identificar una probabilidad de que un suministro de audio contenga una señal acústica que se origine desde la dirección de la fuente de interés, y puede por lo tanto limitar el procesamiento del suministro de audio en base a esa probabilidad. Limitar el procesamiento del suministro de audio de esta manera permite, por ejemplo, que se pueda utilizar detección de actividad de voz de más baja potencia para reducir la cantidad de potencia consumida mientras un dispositivo está operando, por ejemplo, en un modo de siempre en escucha. Los beneficios adicionales de las realizaciones descritas se discuten a través de la descripción.

La Figura 1 es un diagrama de bloques de un entorno 100 operativo en el que pueden ser empleadas varias realizaciones de la presente descripción. Según se ha representado, el entorno 100 operativo incluye un dispositivo 102 de computación. El dispositivo 102 de computación incluye un sistema 104 de procesamiento de sonido. El sistema 104 de procesamiento de sonido puede estar configurado para identificar sonido procedente de una fuente de interés (p. ej., el punto de interés 110). Según se utiliza en la presente memoria, una fuente de interés es una entidad (p. ej., un usuario) que produce, directa o indirectamente, un sonido de interés (p. ej., la voz del usuario), mientras que un punto de interés puede ser utilizado en general para indicar una ubicación, o una ubicación esperada, de una fuente de interés. Se apreciará que, aunque el sistema 104 de procesamiento de sonido es el único componente representado en el dispositivo 102 de computación, esto es únicamente por simplicidad de la explicación. El dispositivo 102 de computación puede contener, o incluir, un número cualquiera de otros componentes que podrán ser fácilmente reconocidos en el estado de la técnica.

Para realizar la identificación de sonido procedente de una fuente de interés, el sistema 104 de procesamiento de sonido, en la realización representada, incluye un primer dispositivo 106 de captura de audio y un segundo dispositivo 108 de captura de audio. Los dispositivos 106 y 108 de captura de audio pueden representar cualquier

tipo de dispositivo, o dispositivos, configurados para capturar sonido, tal como, por ejemplo, un micrófono. Dicho micrófono podría ser de naturaleza omnidireccional o direccional. Los dispositivos 106 y 108 de captura de audio pueden estar configurados para capturar señales acústicas que viajan a través del aire y convertir esas señales acústicas en señales eléctricas. Según se usa en la presente memoria, la referencia a un suministro de audio puede referirse tanto a las señales acústicas capturadas por un dispositivo de captura de audio como a señales eléctricas que son producidas por un dispositivo de captura de audio. Adicionalmente, los dispositivos 106 y 108 de captura de audio pueden ser del mismo tipo de dispositivo de captura de audio o podrán ser diferentes cada uno del otro. Por ejemplo, el dispositivo 106 de captura de audio podría ser un micrófono direccional configurado para un intervalo de respuesta de frecuencia determinado y el dispositivo 108 de captura de audio podría ser un micrófono omnidireccional configurado con el mismo intervalo de respuesta de frecuencia, o con un intervalo de respuesta de frecuencia diferente. Según se ha representado el dispositivo 106 de captura de audio está localizado en relación de proximidad más cercana al punto de interés 110 que el dispositivo 108 de captura de audio. En algunas realizaciones, por ejemplo, cuando una fuente de ruido de fondo es conocida, el dispositivo 108 de captura de audio puede estar situado en relación de proximidad más cercana a una fuente 112 de ruido de fondo. Como tal, se puede suponer, al menos con respecto a la realización representada, que el punto de interés 110 está situado en una posición relativamente coherente alejada del dispositivo 106 de captura de audio para mantener la relación de cercanía mencionada con anterioridad. Adicionalmente, se apreciará que, dependiendo de diversos factores, tal como, por ejemplo, la sensibilidad y la direccionalidad de los respectivos dispositivos de captura de audio, el punto de interés 110 puede necesitar estar situado en una dirección o intervalo de direcciones específicas desde el dispositivo 106 de captura de audio. Por ejemplo, si el dispositivo 106 de captura de audio es un micrófono direccional, entonces la direccionalidad dentro de la cual puede estar localizado el punto de interés 110, puede estar más limitada que si el dispositivo 106 de captura de audio es un micrófono omnidireccional.

El sistema 104 de procesamiento de sonido incluye también un módulo 114 de detección de actividad de voz acoplado con dispositivos 106 y 108 de captura de audio. El módulo 114 de detección de actividad de voz puede estar configurado para recibir y procesar señales, o suministros de audio, presentados a la salida por dispositivos 106 y 108 de captura de audio. Este procesamiento puede permitir que el módulo 114 de detección de actividad de voz identifique sonido que se origina a partir del punto de interés 110, según se discute con detalle más adelante. Se apreciará que, aunque se ha representado en la Figura 1 un módulo 114 de detección de actividad de voz, esta descripción no se limita únicamente a la detección de actividad de voz. El módulo 114 de detección de actividad de voz está destinado simplemente a ser ilustrativo de una implementación posible de la presente descripción y cualquier dispositivo que esté configurado para identificar sonido que se origine a partir de un punto de interés está contemplado explícitamente dentro del alcance de la presente descripción.

Según se ha representado, el módulo 114 de detección de actividad de voz está configurado para recibir un primer suministro de audio desde el dispositivo 106 de captura de audio, y un segundo suministro de audio desde el segundo dispositivo 108 de captura de audio. En realizaciones, el módulo 114 de detección de actividad de voz puede estar configurado para procesar el primer suministro de audio, utilizando el segundo suministro de audio, para permitir la identificación de sonido que se origine desde el punto de interés 110, o sonido que se origine desde la dirección del punto de interés 110.

En algunas realizaciones, el procesamiento del primer suministro de audio utilizando el segundo suministro de audio puede incluir atenuar, o filtrar, frecuencias del primer suministro de audio, que estén compartidas entre el primer suministro de audio y el segundo suministro de audio. Según se usa en la presente memoria, una frecuencia que sea compartida entre dos suministros de audio se refiere a una frecuencia que está contenida dentro de ambos suministros de audio. Por decirlo de otra manera, una frecuencia compartida entre el primer suministro de audio y el segundo suministro de audio podría incluir frecuencias que estén contenidas dentro del primer suministro de audio que también estén contenidas dentro del segundo suministro de audio. La señal de salida de este procesamiento puede ser un suministro de audio atenuado, o filtrado. Atenuar las frecuencias del primer suministro de audio que existan dentro del segundo suministro de audio incluye reducir la amplitud de esas frecuencias dentro del primer suministro de audio. Por el contrario, filtrar frecuencias del primer suministro de audio que existan dentro del segundo suministro de audio incluye eliminar esas frecuencias compartidas del primer suministro de audio. En algunas realizaciones, dicho filtraje puede también tener en cuenta las amplitudes de las frecuencias respectivas. En dichas realizaciones, las frecuencias que son filtradas del primer suministro de audio podrían solamente ser eliminadas en la medida de la amplitud de la frecuencia contenida dentro del segundo suministro de audio. Por ejemplo, si una frecuencia compartida tiene una amplitud X en el primer suministro de audio y una amplitud Y en el segundo suministro de audio, la frecuencia filtrada resultante puede tener una amplitud de X-Y. Si Y es mayor que X, entonces la frecuencia filtrada resultante solamente puede ser completamente eliminada del primer suministro de audio. Este procesamiento ha sido representado, y además discutido, con referencia a la Figura 2A que sigue.

Para realizar el procesamiento que antecede del primer suministro de audio utilizando el segundo suministro de audio, el primer suministro de audio y el segundo suministro de audio pueden necesitar estar sincronizados en el tiempo cada uno con el otro. Según se usa en la presente memoria, sincronizar en el tiempo dos suministros de audio se refiere a alinear los dos suministros de audio en un instante del tiempo de tal modo que los dos suministros de audio puedan ser comparados uno frente al otro en un instante del tiempo. Por ejemplo, el sonido producido pro el punto de interés 110 alcanzará el dispositivo 106 de captura de audio con anterioridad a alcanzar el dispositivo 108 de captura de audio. Como tal, sincronizar en el tiempo el primer suministro de audio con el segundo suministro

de audio podría incluir aplicar un retardo al primer suministro de audio para tener en cuenta el retardo entre el sonido que alcanza el dispositivo 106 de captura de audio y el mismo sonido que alcanza el dispositivo 108 de captura de audio. En consecuencia, en dicho ejemplo, el retardo aplicado al primer suministro de audio podría representar la cantidad de tiempo que se necesita para que el sonido viaje desde el dispositivo 106 de captura de audio hasta el dispositivo 108 de captura de audio.

En varias realizaciones, el módulo 114 de detección de actividad de voz puede también estar configurado para procesar el segundo suministro de audio, utilizando el primer suministro de audio, para facilitar aún más la identificación del sonido que se origina desde el punto de interés 110, o al menos el sonido que se origina desde la dirección del punto de interés 110. En dichas realizaciones, el procesamiento del segundo suministro de audio utilizando el primer suministro de audio puede ser imagen espejo del procesamiento del primer suministro de audio utilizando el segundo suministro de audio que se ha discutido con anterioridad. Por ejemplo, este procesamiento podría incluir atenuar, o filtrar, frecuencias del segundo suministro de audio, que son compartidas entre el segundo suministro de audio y el primer suministro de audio. La señal de salida de este procesamiento puede ser otro suministro de audio, atenuado o filtrado. Este procesamiento ha sido representado, y se discute además, con referencia a la Figura 2B en lo que sigue.

Al igual que con el procesamiento del primer suministro de audio, realizar el procesamiento que antecede del segundo suministro de audio utilizando el primer suministro de audio puede incluir sincronizar en el tiempo el segundo suministro de audio con el primer suministro de audio. Esta sincronización en el tiempo podría ser la imagen de espejo de lo que se ha discutido con anterioridad haciendo referencia a la sincronización en el tiempo del primer suministro de audio con el segundo suministro de audio. Por ejemplo, el sonido producido por el ruido 112 de fondo alcanzará el dispositivo 108 de captura de audio con anterioridad a alcanzar el dispositivo 106 de captura de audio. Como tal, sincronizar en el tiempo el segundo suministro de audio con el primer suministro de audio podría incluir aplicar un retardo al segundo suministro de audio para tomar en consideración el retardo entre el sonido que alcanza el dispositivo 108 de captura de audio y el mismo sonido que alcanza el dispositivo 106 de captura de audio. Por consiguiente, en dicho ejemplo, el retardo aplicado al primer suministro de audio podrá representar la cantidad de tiempo que se necesita para que el sonido viaje desde el dispositivo 106 de captura de audio hasta el dispositivo 108 de captura de audio.

El módulo 114 de detección de actividad de voz puede entonces, en algunas realizaciones, estar configurado para comparar varias bandas de frecuencia, o intervalos de frecuencia, entre el suministro de audio atenuado, o filtrado, producido a partir del primer suministro de audio, mencionado simplemente en lo que sigue como primer suministro de audio procesado, y el suministro de audio atenuado, o filtrado, producido a partir del segundo suministro de audio, mencionado simplemente en lo que sigue como segundo suministro de audio procesado. El módulo 114 de detección de actividad de voz puede estar configurado para determinar un nivel de confianza de fuente que sea indicativo de si el sonido se está originando desde el punto de interés 110. Dicha determinación puede estar basada en el número de bandas de frecuencia del primer suministro de audio procesado que exceda un umbral predefinido, o preconfigurado, de diferencia con bandas correspondientes del segundo suministro de audio procesado. En realizaciones, un valor más alto para el nivel de confianza de fuente puede ser más indicativo de sonido dentro del primer suministro de audio procesado que se origina desde el punto de interés 110 que un valor más bajo para el nivel de confianza de fuente.

En varias realizaciones, el módulo 114 de detección de actividad de voz puede estar también configurado para comparar las diversas bandas de frecuencia mencionadas con anterioridad, o intervalos de frecuencia, entre el primer suministro de audio procesado y el segundo suministro de audio procesado para determinar un nivel de confianza del ruido, o de ruido de fondo. Este nivel de confianza de ruido es indicativo de si el primer suministro de audio procesado es ruido. Dicha determinación puede estar basada en el número de bandas de frecuencia del primer suministro de audio procesado que están dentro de un umbral predefinido, o preconfigurado, de diferencia respecto a las bandas de frecuencia correspondientes del segundo suministro de audio procesado. En realizaciones, un valor más alto del nivel de confianza de ruido puede ser más indicativo de que el sonido sea ruido dentro del primer suministro de audio procesado que un valor más bajo para el nivel de confianza de ruido.

Se apreciará que, mientras que la descripción anterior está dirigida hacia una realización en que el punto de interés 110 está situado en relación de proximidad más cercana al dispositivo 106 de captura de audio, la localización del punto de interés 110 podría cambiar de tal modo que el punto de interés se sitúe en relación de proximidad más cercana al dispositivo 108 de captura de audio. En ese escenario, el módulo 114 de detección de actividad de voz puede estar configurado para conmutar el procesamiento descrito con anterioridad de tal modo que el suministro de audio capturado por el dispositivo 108 de captura de audio sea procesado para identificar audio que se origina a partir del punto de interés recientemente localizado. En diversas realizaciones, la conmutación podría ser realizada programáticamente (p. ej., a través de lógica codificada en el módulo 114 de detección de actividad de voz) o en la selección por un usuario del dispositivo 102 de computación (p. ej., a través de una interfaz de usuario, un comando de voz o una conmutación de hardware).

Según se ha representado, en algunas realizaciones, el sistema 104 de procesamiento de sonido incluye también un módulo 116 de cancelación de eco acústico (AEC). En tales realizaciones, el módulo 114 de detección de actividad de voz puede presentar a la salida un suministro de audio para el módulo 116 de AEC. El suministro de audio de

salida podría ser, por ejemplo, el primer suministro de audio procesado, o el primer suministro de audio en sí mismo, dado que estos suministros de audio podrían incluir una amplitud más alta para esos sonidos, o frecuencias, que se originen desde la dirección del punto de interés 110. El módulo 116 de AEC puede estar configurado para reducir una cantidad de eco contenida dentro del suministro de audio proporcionado a la salida por medio del módulo 114 de detección de actividad de voz. Dichas configuraciones de AEC son conocidas en el estado de la técnica y no van a ser descritas aquí con mayor detalle.

En algunas realizaciones, que el módulo 114 de detección de actividad de voz presente a la salida un suministro de audio para el modulo 116 de AEC, podría estar supeditado a que el nivel de confianza de fuente del primer suministro de audio procesado alcance, o exceda, un umbral de confianza de fuente, o límite. En otras realizaciones, que el módulo 114 de detección de actividad de voz presente a la salida un suministro de audio para el módulo 116 de AEC podría estar supeditado a que el nivel de confianza de ruido del primer suministro de audio procesado alcance, o supere, un umbral de confianza de ruido, o límite. Como tal, el módulo 114 de detección de actividad de voz podría limitar aquellos casos en que se presente a la salida un suministro de audio para aquellos casos en que el módulo de detección de actividad de voz haya establecido un nivel de confianza suficiente de que el suministro de audio incluye sonido que se ha originado desde la dirección del punto de interés para justificar el procesamiento adicional. Al hacer eso, el módulo 114 de detección de actividad de voz puede reducir la energía consumida por el módulo 116 de AEC, así como cualquier procesamiento posterior (p. ej., por medio del modulo 118 de reconocimiento de voz), y ahorrar con ello energía del dispositivo 102 de computación, reduciendo la cantidad del suministro de audio de salida que se procesa adicionalmente.

El umbral de confianza de fuente o el lumbral de confianza de ruido podrán estar predefinidos, preconfigurados, o podrán ser determinados programáticamente. En algunas realizaciones, el umbral de confianza de fuente, o el umbral de confianza de ruido, podrían estar basados en un nivel de potencia actual del dispositivo 102 de computación. Por ejemplo, si el dispositivo 102 de computación está operando con una batería llena, o está actualmente enchufado a una fuente de potencia continua, el umbral de confianza de fuente podría ser establecido en un valor más bajo que si la batería del dispositivo 102 de computación está operando a un nivel de potencia más bajo. Como tal, el umbral de confianza de fuente puede ser ajustado, en algunas realizaciones, en un valor más alto que lo que desciende el nivel de potencia del dispositivo 102 de computación en un esfuerzo por ahorrar más vida de la batería limitando la cantidad de suministro de audio que es procesado por el módulo 116 de AEC, y a continuación por cualesquiera otros módulos.

El sistema 104 de procesamiento de sonido puede incluir también opcionalmente un módulo 118 de reconocimiento de voz. El módulo 118 de reconocimiento de voz podría estar configurado para monitorizar el suministro de audio recibido por el módulo 118 de reconocimiento voz, para identificar uno o más activadores contenidos dentro del suministro de audio recibido. El suministro de audio recibido por el módulo 118 de reconocimiento de voz podrá proceder del módulo 116 de AEC, en realizaciones en las que el módulo 116 de AEC haya sido incluido. En otras realizaciones, donde el módulo 116 de AEC no haya sido incluido en el sistema 114 de procesamiento de sonido, o esté incluido con anterioridad al módulo 114 de detección de actividad de voz, el módulo 118 de reconocimiento de voz podrá recibir el suministro de audio directamente desde el módulo 114 de detección de actividad de voz. En tales realizaciones, el módulo 114 de detección de actividad de voz podrá estar configurado, según se ha discutido con anterioridad con referencia al módulo 116 de AEC, de modo que solamente presente a la salida un suministro de audio para el módulo 118 de reconocimiento de voz cuando el módulo 114 de detección de actividad de voz haya establecido un nivel de certidumbre suficiente de que el suministro de audio incluye audio que se origina desde la dirección del punto de interés. Esto puede ser especialmente ventajoso en escenarios en que el dispositivo 102 de computación esté capacitado paras operar en un modo de siempre en escucha. Según se utiliza en la presente memoria, un modo de siempre en escucha es uno en que el sistema 104 de procesamiento de sonido está configurado para capturar y procesar continuamente audio que identifique los activadores contenidos dentro del audio. Ejemplos de aplicaciones que pueden utilizar un modo de siempre en escucha están representados por Cortana ofrecida por Microsoft Corp., de Redmond, Washington; Google Now ofrecido por Google Co., de Mountain View, o Sin, ofrecido por Apple Inc. de Cupertino, California.

Según se ha mencionado con anterioridad, el suministro de audio capturado por el dispositivo 106 de captura de audio podrá incluir una amplitud más alta para aquellos sonidos, o frecuencias, que se originen desde la dirección del punto de interés 110, y por lo tanto, el primer suministro de audio o una versión procesada del primer suministro de audio (p. ej., filtrado, atenuado o procesado por el módulo 116 de AEC) podrá ser proporcionado al módulo 118 de reconocimiento de voz para identificar activadores que se originen desde el punto de interés 110.

Un problema que se presenta normalmente con los modos de siempre en escucha mencionados con anterioridad, consiste en limitar el procesamiento del suministro de audio a aquellos casos en los que el suministro de audio se origina desde el punto de interés 110 (p. ej., un usuario). Limitando el procesamiento de los suministros de audio a suministros de audio que incluyan señales acústicas que se originen desde el punto de interés, según se ha descrito con anterioridad, la cantidad de procesamiento requerido para operar en el modo de siempre en escucha se reduce, lo que reduce correspondientemente la cantidad de energía necesaria para operar en el modo de siempre en escucha. Otro problema que se encuentra con el modo de siempre en escucha consiste en la capacidad para activar una acción que no fue iniciada por el usuario. Por ejemplo, una persona nefasta podría pasar caminando y dar una orden (p. ej., un comando de apagado, un comando de encendido, etc.) al dispositivo 102 de computación que

provoque que el dispositivo 102 de computación realice una acción no deseada por el usuario. Limitando el procesamiento de suministros de audio a aquellos suministros de audio que incluyen una señal acústica que se origine desde una dirección del punto de interés, según se ha descrito con anterioridad, la capacidad para que una persona nefasta emita dicho comando desde otras direcciones quedaría limitada. Se apreciará que esto se debe a que un usuario nefasto que intente emitir un comando de ese tipo desde otra dirección tendría ese comando que alcanzar el dispositivo de captura de audio (p. ej., el dispositivo 108 de captura de audio) que está situado más allá del punto de interés. Como resultado, la amplitud para ese comando del usuario nefasto sería más alta en el suministro de audio capturado por el dispositivo de captura de audio más allá del punto de interés, y más baja en el suministro de audio capturado por el dispositivo de captura de audio que está en relación de proximidad más cercana al punto de interés.

Se apreciará que los beneficios de las realizaciones descritas con anterioridad pueden extenderse más allá de un modo de siempre en escucha. Por ejemplo, el umbral de confianza de ruido descrito con anterioridad podría ser utilizado para identificar de manera más eficiente ruido de fondo. Como tal, cualesquiera aplicaciones que necesiten identificar ruido de forma precisa podrían beneficiarse también de las realizaciones descritas con anterioridad. Por ejemplo, los codificadores de habla codifican con frecuencia ruido identificado con un número más bajo de bits que el habla. Esto permite una tasa media de bits más baja para un suministro de audio, lo que puede reducir la cantidad de procesamiento del suministro de audio, reduciendo con ello el consumo de potencia de un dispositivo de computación que realice el procesamiento. Adicionalmente, las aplicaciones de reducción de ruido que buscan estimar de forma precisa características de ruido de un entorno, podrían beneficiarse también de las realizaciones descritas con anterioridad, en particular, aquellas que incluyen el umbral de confianza de ruido. Los beneficios y las aplicaciones adicionales de las realizaciones descritas con anterioridad podrán ser fácilmente comprendidos por los expertos en la materia, y los ejemplos anteriores están destinados simplemente a ilustrar una muestra de los beneficios que las realizaciones descritas con anterioridad pueden proporcionar.

Las Figuras 2A, 2B y 2C muestran representaciones esquemáticas ilustrativas de configuraciones de sistema de procesamiento de sonido, conforme a varias realizaciones de la presente descripción. La Figura 2A muestra una representación ilustrativa de una porción de un sistema 202 de procesamiento de sonido configurado para procesar dos suministros de audio, tales como los que se han discutido con referencia a la Figura 1. Según se muestra, el sistema 202 de procesamiento de sonido incluye micrófonos 206 y 208. Tal y como puede apreciarse, el micrófono 206 está ubicado en relación de proximidad más cercana a una fuente de interés 204 que el micrófono 208, y los micrófonos 206 y 208 están situados a una distancia ‘d’ cada uno del otro.

El micrófono 206 puede estar configurado para que capture un primer suministro de audio, representado por $X_1(\omega,\theta)$ 210, mencionado en lo que sigue simplemente como “primer suministro 210 de audio”, donde ω representa cada frecuencia, o intervalo de frecuencia, contenida dentro del primer suministro 210 de audio. El micrófono 208 puede estar configurado para capturar un segundo suministro de audio, representado en la presente por $X_2(\omega,\theta)$ 212, mencionado en lo que sigue simplemente como “segundo suministro 212 de audio”. Para procesar los dos suministros de audio puede ser necesario sincronizar en el tiempo el segundo suministro 210 de audio con el primer suministro 212 de audio. Tal sincronización en el tiempo se ha discutido con detalle mediante referencia a la Figura 1, en lo que antecede, y puede incluir aplicar un retardo al segundo suministro 212 de audio. Este retardo está representado por τ_1 en la casilla 214, mencionado en lo que sigue simplemente como retardo 214. El retardo 214 puede reflejar la cantidad de tiempo que se necesita para que el sonido viaje desde el primer micrófono 206 hasta el segundo micrófono 208 a través de la distancia ‘d’.

El primer y el segundo suministros de audio sincronizados en el tiempo pueden ser recibidos en 216, donde, según se ha indicado mediante los operadores adyacentes a los respectivos suministros de audio, el primer suministro de audio se atenúa, o se filtra, utilizando el segundo suministro de audio para producir un suministro de audio atenuado, o filtrado, representado por $C_B(\omega,\theta)$ 218, mencionado en lo que sigue simplemente como suministro 218 de audio procesado. De nuevo, ω representa cada frecuencia, o intervalo de frecuencia, contenida dentro del suministro 218 de audio procesado. Los expertos en la materia podrán apreciar que $C_B(\omega,\theta)$ representa una cardioide de audio que está representada por el suministro 218 de audio procesado. También se apreciará que la representación mostrada puede ser relacionada con el estado de la técnica disponiendo un valor nulo a 0 grados.

La Figura 2B muestra una representación ilustrativa de otra porción de un sistema 222 de procesamiento de sonido configurado para procesar el primer suministro 210 de audio discutido anteriormente y un segundo suministro 212 de audio; sin embargo, tal y como puede apreciarse, la configuración representada es una imagen de espejo de la que se ha discutido con anterioridad con referencia a la Figura 2A. Como tal, la porción de sistema 222 de procesamiento de sonido muestra el procesamiento del segundo suministro 212 de audio utilizando el primer suministro 212 de audio. Para realizar este procesamiento puede ser necesario sincronizar en el tiempo el primer suministro 210 de audio y el segundo suministro 212 de audio. Según se ha mencionado con anterioridad, la sincronización en el tiempo puede incluir aplicar un retardo al primer suministro 212 de audio. Este retardo está representado por τ_2 en la casilla 224, siendo mencionado simplemente en lo que sigue como retardo 224. El retardo 224 puede reflejar la cantidad de tiempo que se necesita para que el sonido viaje desde el primer micrófono 206 hasta el segundo micrófono 208 a través de la distancia ‘d’.

El primer y el segundo suministros de audio sincronizados en el tiempo pueden ser recibidos en 226, donde, según se ha indicado mediante los operadores adyacentes a los respectivos suministros de audio, el segundo suministro de audio se atenúa, o se filtra, utilizando el primer suministro de audio para producir un suministro de audio atenuado, o filtrado, representado aquí por $C_F(\omega,\theta)$ 228, mencionado en lo que sigue simplemente como suministro 228 de audio procesado. De nuevo, ω representa cada frecuencia, o intervalo de frecuencia, contenida dentro del suministro 228 de audio procesado. Los expertos en la materia podrán apreciar que $C_F(\omega,\theta)$ representa una cardioide de audio que está representada por el suministro 228 de audio procesado. También se apreciará que la representación mostrada puede ser referida al estado de la técnica colocando un valor nulo a 180 grado.

La Figura 2C muestra una representación ilustrativa de las porciones de sistema 202 y 222 de sonido, discutidas con anterioridad, combinadas en un único sistema. Como tal, cada uno de los aspectos de las Figuras 2A y 2B discutidos con anterioridad, han sido representados en la Figura 2C.

Las Figuras 3A y 3B son representaciones gráficas de niveles de confianza de fuente y niveles de confianza de ruido, conforme a varias realizaciones de la presente descripción. La Figura 3A es una representación ilustrativa de un ejemplo de nivel de confianza de fuente. Según se puede apreciar, el cálculo para determinar el nivel de confianza de fuente representado la Figura 3A se basa en un ejemplo de algoritmo definido por $C_F(\omega) - C_B(\omega) > \Delta_1(\omega) \rightarrow Cnt_1{+}{+}$, donde $C_F(\omega)$ representa una frecuencia, o una banda de frecuencia, ω dentro de una cardiode frontal, mencionado también en la presente memoria como suministro de audio procesado (p. ej., el suministro 218 de audio procesado, de las Figuras 2A y 2C); $C_B(\omega)$ representa la misma frecuencia, o banda de frecuencia, ω dentro de una cardiode posterior, también mencionada en la presente memoria como suministro de audio procesado (p. ej., el suministro 228 de audio procesado de las Figuras 2B y 2C); $\Delta_1(\omega)$ representa un umbral predefinido de diferencia, y $Cnt_1{+}{+}$ representa un recuento actualizado de esas frecuencias, o bandas de frecuencia, que exceden del umbral de diferencia, $\Delta_1(\omega)$. El gráfico 300 representa el recuento actualizado, Cnt_1, a lo largo del eje X y un nivel de confianza de fuente, P_v, a lo largo del eje Y. Tal y como puede apreciarse, según se incrementa el recuento actualizado de frecuencias que exceden el umbral de diferencia entre la cardiode frontal y la cardioide posterior, lo hace también el nivel de confianza de fuente. Según se ha representado, la línea 306 punteada representa una función que significa un límite de confianza de fuente, mencionado en lo que sigue como “función 306 de límite de confianza de fuente”, más allá del cual el nivel de confianza de fuente tiene suficientemente establecido que la cardioide frontal incluye audio que se origina desde la fuente de interés, o desde la dirección de la fuente de interés. En realizaciones, si el nivel de confianza de fuente ha sido suficientemente establecido, entonces se puede permitir el procesamiento adicional de la cardiode frontal, o del suministro de audio que fue procesado (p. ej., atenuado o filtrado) para producir la cardioide frontal (p. ej., a través de reconocimiento de voz). Como tal, un nivel de confianza de fuente que esté por debajo de la línea 310 no podría ser establecido suficientemente, y no se permitiría pasar a través del mismo para el procesamiento adicional. En concordancia con la función 306 de límite de confianza de fuente, se puede ver que un valor Cnt_1 de 308 podría coincidir con un nivel de confianza de fuente suficiente. Se apreciará que esto está destinado simplemente a ilustrar una posible determinación de nivel de confianza de fuente. Según se ha mencionado con anterioridad, la función 306 de límite de confianza de fuente puede ser ajustada dependiendo de los detalles de implementación o dependiendo de un estado actual (p. ej., nivel de batería) del dispositivo de computación que esté implementando dicho límite de confianza de fuente. Adicionalmente, se apreciará que en la materia se pueden utilizar otros métodos, o algoritmos, para determinar un nivel de confianza de fuente sin apartarse del alcance de la presente descripción.

La Figura 3B, por el contrario, es una representación ilustrativa de un ejemplo de nivel de confianza de ruido. El nivel de confianza de ruido representado en la Figura 3B se basa en un ejemplo de algoritmo definido por $|C_F(\omega) - C_B(\omega)| < \Delta_2(\omega) \rightarrow Cnt_2{+}{+}$, donde de nuevo $C_F(\omega)$ representa una frecuencia, o un intervalo de frecuencia, ω dentro de una cardioide frontal; $C_B(\omega)$ representa la misma frecuencia, o banda de frecuencia, ω dentro de una cardioide posterior; $\Delta_2(\omega)$ representa un umbral predefinido de diferencia, y $Cnt_2{+}{+}$ representa un recuento actualizado de esas frecuencias, o bandas de frecuencia, que están dentro de un umbral de diferencia, $\Delta_2(\omega)$. El gráfico 320 representa el recuento actualizado, Cnt_2, a lo largo del eje X y un nivel de confianza de ruido, Pd, a lo largo del eje Y. Tal y como puede apreciarse, según se incrementa el recuento actualizado de frecuencias que están dentro del umbral de diferencia entre la cardioide frontal y la cardioide posterior, también lo hace el nivel de confianza de ruido. Tal y como se ha representado, la línea 314 de puntos representa una función que significa un límite de confianza de ruido, mencionado en lo que sigue como “función 314 de límite de confianza de ruido”, más allá del cual el nivel de confianza de ruido ha establecido suficientemente que la cardioide frontal incluye ruido (p. ej., ruido de fondo) en vez de audio que se origine a partir de la fuente de interés, o de la dirección de la fuente de interés. En realizaciones, si el nivel de confianza de ruido ha sido suficientemente establecido, entonces puede que no se permita el procesamiento adicional de la cardioide frontal, o del suministro de audio que fue procesado (p. ej., atenuado o filtrado) para producir la cardioide frontal. Como tal, un nivel de confianza de ruido que esté por debajo de la línea 318 no podrá ser suficientemente establecido y se podrá permitir que pase a su través para un procesamiento adicional. Conforme a la función 314 de límite de confianza de ruido, se puede apreciar que un valor Cnt_2 de 316 podría coincidir con un nivel de confianza de fuente suficiente. Se apreciará que esto está destinado simplemente a ilustrar una posible determinación de nivel de confianza de rudo. Según se ha mencionado con anterioridad, la función 314 de límite de confianza de ruido puede ser ajustada dependiendo de los detalles de implementación o dependiendo de un estado actual (p. ej., nivel de batería) del dispositivo de computación que esté implementando

dicho límite de confianza de ruido. Adicionalmente, se apreciará en el estado de la técnica que otros métodos, o algoritmos, para determinar un nivel de confianza de ruido pueden ser utilizados sin apartarse por ello del alcance de la presente descripción.

La Figura 4 muestra una representación esquemática ilustrativa de un sistema 400 de procesamiento de sonido que tiene una configuración de tres micrófonos, conforme a varias realizaciones de la presente divulgación. Por motivos de claridad, varios aspectos del sistema de procesamiento de sonido han sido agrupados en bloques 401a y 401b. Estos bloques se utilizan simplemente a efectos de referencia para distribuir la funcionalidad del sistema de procesamiento de sonido en unidades similares a las representadas en la Figura 2C y que no se deberá considerar como limitativo en ningún aspecto de la presente descripción. Según se ha representado, el sistema 400 de procesamiento de sonido incluye micrófonos 402, 404 y 406. Cada una de las fuentes 408-414 representan posibles fuentes de sonido y cualquiera de las fuentes 408-414 podría ser una fuente de interés. Como tal, uno cualquiera de los micrófonos 402-406 podría estar situado en relación de proximidad más cercana a una fuente de interés que los otros dos micrófonos.

El micrófono 402 puede estar configurado para capturar un primer suministro de audio, representado en la presente memoria por $X_1(\omega,\theta)$ 416, mencionado en lo que sigue simplemente como “primer suministro 416 de audio”, donde ω representa cada frecuencia, o intervalo de frecuencia, contenida dentro del primer suministro 416 de audio. El micrófono 404 puede estar configurado para capturar un segundo suministro de audio, representado en este caso por $X_2(\omega,\theta)$ 418, mencionado en lo que sigue simplemente como “segundo suministro 418 de audio”. El micrófono 406 puede estar configurado para capturar un tercer suministro de audio, representado en este caso por $X_3(\omega,\theta)$ 420, mencionado en lo que sigue simplemente como “tercer suministro 420 de audio”.

Tal y como puede apreciarse, los suministros 416-420 de audio son procesados por pares, siendo el segundo suministro 418 de audio procesado dos veces, según se ha indicado por medio de las cuatro flechas que salen del micrófono 404, una vez dentro del bloque 401a con el suministro 416 de audio y una vez dentro del bloque 401b con el suministro 420 de audio.

Empezando por el bloque 401a, para procesar el primer suministro 416 de audio y el segundo suministro 418 de audio, los dos suministros de audio pueden necesitar ser sincronizados en el tiempo, según se discute en otras posiciones de la presente memoria. Según se ha representado dicha sincronización en el tiempo puede incluir aplicar un retardo (p. ej., 422a-422b) al respectivo suministro de audio que está siendo utilizado para procesar (p. ej., filtrar, atenuar, etc.) el otro suministro de audio. Por ejemplo en 424a, se está utilizando el primer suministro 416 de audio para procesar el segundo suministro 418 de audio, según se ha indicado por medio de los operadores adyacentes a los respectivos suministros de audio, para producir un suministro de audio procesado representado por $C_{F1}(\omega,\theta)$ 426a, mencionado simplemente en lo que sigue como suministro 426a de audio procesado. Como resultado, el primer suministro 416 de audio ha tenido un retardo 422a aplicado al mismo. Adicionalmente, en 424b, el segundo suministro 418 de audio está siendo utilizado para procesar el primer suministro 416 de audio, según se ha indicado por medio de los operadores adyacentes a los respectivos suministros de audio, para producir un suministro de audio procesado representado por $C_{B1}(\omega,\theta)$ 426b, mencionado simplemente en lo que sigue como suministro 426b de audio procesado. Como resultado, el segundo suministro 418 de audio ha tenido un retardo 422b aplicado al mismo. Los retardos 422a y 422b pueden reflejar la cantidad de tiempo que se necesita para que el sonido viaje entre el micrófono 402 y el micrófono 404. Se apreciará que, en algunas realizaciones, se podría invertir el procesamiento en 424a y 424b de tal modo que el retardo sea aplicado al suministro de audio que se está procesando. En una realización de ese tipo, 424a podría presentar a la salida $C_{F1}(\omega,\theta)$ y 424b podría presentar a la salida $C_{B1}(\omega,\theta)$.

Siguiendo hasta el bloque 401b, para procesar el segundo suministro 418 de audio y el tercer suministro 420 de audio los dos suministros podrían necesitar también estar sincronizados en el tiempo. Según se ha representado, esa sincronización en el tiempo puede incluir aplicar un retardo (p. ej., 422c-422d) al respectivo suministro de audio que esté siendo utilizado para procesar (p. ej., filtrar, atenuar, etc.) el otro suministro de audio. Por ejemplo, en 424c, el segundo suministro 416 de audio está siendo utilizado para procesar el tercer suministro 418 de audio, según se indica por medio de los operadores adyacentes a los respectivos suministros de audio recibidos por 424c, para producir un suministro de audio procesado representado por $C_{F2}(\omega, \theta)$ 426c, mencionado simplemente en lo que sigue como suministro 426c de audio procesado. Como resultado el segundo suministro 418 de audio ha tenido un retardo 422c aplicado al mismo. Además, en 424d, el tercer suministro 420 de audio está siendo utilizado para procesar el segundo suministro 418 de audio, según se ha indicado mediante los operadores adyacentes a los respectivos suministros de audio recibidos en 424d, para producir un suministro de audio procesado representado por $C_{B2}(\omega, \theta)$ 426d, mencionado simplemente en lo que sigue como suministro 426d de audio procesado. Como resultado, el tercer suministro 420 de audio ha tenido un retardo 422d aplicado al mismo. Los retardos 422c y 422d reflejan la cantidad de tiempo que se necesita para que el sonido viaje entre el micrófono 404 y el micrófono 406. Al igual que con 424a y 424b, se apreciará que, en algunas realizaciones, el procesamiento en 424c y 424d podría ser invertido, de tal modo que el retardo sea aplicado al suministro de audio que se está procesando. En una realización de ese tipo, 424c podría presentar a la salida $C_{B2}(\omega, \theta)$ y 424d podría presentar a la salida $C_{F2}(\omega, \theta)$.

La Figura 5 es un diagrama de flujo que representa un método 500 ilustrativo para identificar sonido procedente de una fuente de interés, conforme a varias realizaciones de la presente descripción. El método 500 puede ser llevado a cabo, por ejemplo, mediante un detector de actividad de voz. El método 500 empieza en el bloque 510 donde se recibe un primer suministro de audio capturado por un primer micrófono de un dispositivo de computación. En el bloque 520, se recibe un segundo suministro de audio capturado por un segundo micrófono del dispositivo de computación. Se apreciará que, el bloque 510 y el bloque 520 pueden ocurrir de forma simultánea, al menos de forma sustancialmente simultánea. Según se ha mencionado anteriormente haciendo referencia a la Figura 1, estos micrófonos pueden ser de cualquier tipo, clase o combinación de micrófonos. En realizaciones, el primer micrófono puede estar situado más cerca de un punto de interés que el segundo micrófono. En esas realizaciones, el audio que se origina desde el punto de interés podría ser de mayor magnitud cuando se captura por medio del primer micrófono que cuando se captura por medio del segundo micrófono.

En el bloque 530 se procesan el primer suministro de audio y el segundo suministro de audio para identificar sonido que se origina desde el punto de interés. En algunas realizaciones, este procesamiento puede empezar mediante sincronización en el tiempo del primer suministro de audio con el segundo suministro de audio. Esta sincronización en el tiempo puede ser llevada a cabo, por ejemplo, aplicando un retardo a uno de entre el primer o el segundo suministros de audio, según se ha descrito con anterioridad.

En algunas realizaciones, el procesamiento del primer suministro de audio y del segundo suministro de audio puede incluir procesar el primer suministro de audio utilizando el segundo suministro de audio. En tales realizaciones, el procesamiento puede incluir atenuar, o filtrar, frecuencias del primer suministro de audio, que estén compartidas entre el primer suministro de audio y el segundo suministro de audio, según se ha descrito con referencia a la Figura 1. En varias realizaciones, el procesamiento del primer suministro de audio y del segundo suministro de audio pueden incluir también procesar el segundo suministro de audio, utilizando el primer suministro de audio, para permitir mejor la identificación de sonido que se origine desde el punto de interés, o al menos sonido que se origine desde la dirección del punto de interés. De nuevo, en esas realizaciones, el procesamiento puede incluir atenuar, o filtrar, frecuencias del segundo suministro de audio, que sean compartidas entre el primer suministro de audio y el segundo suministro de audio, según se ha descrito con referencia a la Figura 1.

Otra realización que representa el procesamiento de un primer y un segundo suministros de audio, representado mediante el bloque 530 de la Figura 5, está representado mediante el flujo de proceso 600 de la Figura 6. El flujo de proceso 600 empieza en el bloque 610, donde frecuencias contenidas dentro del primer suministro de audio son atenuadas, o filtradas, en base a frecuencias correspondientes del segundo suministro de audio para producir un primer suministro de audio procesado. En el bloque 620, frecuencias dentro del segundo suministro de audio son atenuadas, o filtradas, en base a frecuencias correspondientes contenidas dentro del primer suministro de audio, para producir un segundo suministro de audio procesado.

En el bloque 630, las bandas de frecuencia contenidas dentro del primer suministro de audio procesado y del segundo suministro de audio procesado se comparan cada una con otra (p. ej., en cuanto a diferencias de amplitud). En el bloque 640, se puede determinar un nivel de confianza de fuente en base a la comparación que se realice en el bloque 630. Este nivel de confianza de fuente es indicativo de si el sonido se está originando desde el punto de interés, o desde la dirección del punto de interés. Esa determinación puede estar basada en el número de bandas de frecuencia del primer suministro de audio procesado que exceda un umbral predefinido, o preconfigurado, de diferencia respecto a las bandas de frecuencia correspondientes del segundo suministro de audio procesado. En realizaciones, un valor más alto para el nivel de confianza de fuente puede ser más indicativo de sonido dentro del primer suministro de audio procesado que se origine desde el punto de interés que un valor más bajo para el nivel de confianza de fuente.

En el bloque 650, se toma una determinación respecto a si el nivel de confianza de fuente, determinado en el bloque 640, excede un límite preconfigurado (p. ej., el límite de confianza de fuente). Según se ha mencionado con anterioridad, este límite preconfigurado puede cambiar dependiendo de un estado (p. ej., nivel de carga) del dispositivo de computación que realiza el flujo de proceso 600. Si el nivel de confianza de fuente no excede el límite preconfigurado, entonces el procesamiento puede volver al bloque 610 y este proceso puede ser repetido. Si, no obstante, el nivel de confianza de fuente excede el límite preconfigurado, entonces el procesamiento avanza hasta el bloque 660 donde el primer suministro de audio, o el primer suministro de audio procesado, se envía a un motor de reconocimiento de voz del dispositivo de computación.

Habiendo descrito brevemente una visión general de las realizaciones de la presente divulgación, a continuación se describe un entorno operativo ilustrativo en el que pueden ser implementadas las realizaciones de la presente descripción con el fin de proporcionar un contexto general para diversos aspectos de la presente descripción. Haciendo inicialmente referencia a la Figura 7 en particular, un entorno operativo ilustrativo para implementar realizaciones de la presente descripción ha sido mostrado y designado en general como dispositivo 700 de computación. El dispositivo 700 de computación es sólo un ejemplo de un entorno de computación adecuado y no se pretende que sugiera ninguna limitación en cuanto al alcance de uso o la funcionalidad de la descripción. En ningún caso debe ser interpretado el dispositivo 700 de computación como que tiene alguna dependencia o requisito relativo con uno cualquiera o con alguna combinación de componentes ilustrados.

La divulgación puede ser descrita en el contexto general de un código informático o de instrucciones utilizables por máquina, incluyendo instrucciones ejecutables con ordenador tal como módulos de programa o motores, que sean ejecutadas por medio de un ordenador u otra máquina, tal como un asistente personal de datos u otro dispositivo portátil. En general, los módulos informáticos que incluyen rutinas, programas, objetos, componentes, estructuras de datos, etc., se refieren a códigos que realizan tareas particulares o que implementan tipos de datos abstractos particulares. La descripción puede ser puesta en práctica mediante una diversidad de configuraciones de sistema, incluyendo los dispositivos portátiles, electrónica de consumo, ordenadores de propósito general, más especialmente dispositivos de computación, etc. La descripción puede también ser puesta en práctica en entornos de computación distribuidos donde se realizan tareas por medio de dispositivos de procesamiento remoto que están enlazados a través de una red de comunicaciones.

Con referencia a la Figura 7, el dispositivo 700 de computación incluye un bus 710 que acopla directa o indirectamente los siguientes dispositivos: memoria 712, uno o más procesadores 714, uno o más componentes 716 de presentación, puertos 718 de entradas/salida, componentes 720 de entrada/salida, y una fuente de alimentación 722 ilustrativa. El bus 710 representa lo que puede ser uno o más buses (tal como un bus de direcciones, un bus de datos, o una combinación de ambos). Aunque los diversos bloques de la Figura 7 han sido representados con líneas claramente delineadas por motivos de claridad, en realidad, tales delineaciones no son tan claras y estas líneas pueden superponerse. Por ejemplo, también se puede considerar que un componente de presentación tal como un dispositivo de visualización, sea un componente de E/S. También, los procesadores tienen generalmente memoria en forma de caché. Se reconoce que ésa es la naturaleza del estado de la técnica, y se reitera que el diagrama de la Figura 7 es simplemente ilustrativo de un ejemplo de dispositivo de computación que puede ser usado junto con una o más realizaciones de la presente divulgación. No se hace distinción entre categorías tales como “estación de trabajo”, “servidor”, “ordenador de sobremesa”, “dispositivo portátil”, etc., puesto que todos ellos están contemplados dentro del alcance de la Figura 7 y de la referencia a “dispositivo de computación”.

El dispositivo 700 de computación incluye típicamente una diversidad de medios legibles con ordenador. Los medios legibles con ordenador pueden ser cualquier medio disponible al que se pueda acceder por medio del dispositivo 700 de computación, e incluye tanto medios volátiles como no volátiles, medios extraíbles y medios no extraíbles. A título de ejemplo, y sin limitación, los medios legibles con ordenador pueden comprender medios de almacenaje en ordenador y medios de comunicación.

Los medios de almacenaje en ordenador incluyen medios volátiles y no volátiles, extraíbles y no extraíbles, implementados mediante cualquier método o tecnología para almacenaje de información tal como instrucciones legibles con ordenador, estructuras de datos, módulos de programa u otros datos. Los medios de almacenaje en ordenador incluyen, aunque sin limitación, memoria RAM, ROM, EEPROM, memoria flash u otra tecnología de memoria, CD-ROM, discos versátiles digitales (DVD) u otro almacenaje en disco óptico, casetes magnéticos, cinta magnética, almacenaje en disco magnético u otros dispositivos de almacenaje magnético, o cualquier otro medio que pueda ser usado para almacenar la información deseada y al que se pueda acceder mediante el dispositivo 100 de computación. Los medios de almacenaje en ordenador excluyen las señales en sí mismas.

Los medios de comunicación materializan típicamente instrucciones legibles con ordenador, estructuras de datos, módulos de programa u otros datos en una señal de datos modulada tal como una onda portadora u otro mecanismo de transporte, e incluye cualquier medio de suministro de información. El término “señal de datos modulada” significa una señal que tiene una o más de sus características establecidas o cambiadas de tal manera que codifica información en la señal. A título de ejemplo, y sin limitación, los medios de comunicación incluyen medios cableados tal como una red cableada o conexión cableada directa, y medios inalámbricos tal como acústicos, RF, infrarrojos y otros medios inalámbricos. Combinaciones de cualesquiera de los anteriores podrán también estar incluidas dentro del alcance de medios legibles con ordenador.

La memoria 712 incluye medios de almacenaje en ordenador en forma de memoria volátil y no volátil. Según se ha representado, la memoria 712 incluye instrucciones 724. Las instrucciones 724, cuando se ejecutan mediante el (los) procesador(es) 714, están configurados para que provoquen que el dispositivo de computación realice cualquiera de las operaciones descritas en la presente memoria, con referencia a las figuras discutidas con anterioridad. La memoria puede ser extraíble, no extraíble, o una combinación de ambas. Dispositivos ilustrativos de hardware incluyen memoria de estado sólido, unidades de disco duro, unidades de disco óptico, etc. El dispositivo 700 de computación incluye uno o más procesadores que leen datos a partir de varias entidades tales como la memoria 712 o componentes 720 de E/S. El (los) componente(s) 716 de presentación presenta(n) indicaciones de datos para un usuario u otro dispositivo. Componentes de presentación ilustrativos incluyen un dispositivo de visualización, altavoz, componente de impresión, componente vibrador, etc.

Puertos 718 de E/S permiten que el dispositivo 700 de computación se acople lógicamente a otros dispositivos incluyendo los componentes 720 de E/S, algunos de los cuales pueden estar integrados. Los componentes ilustrativos incluyen un micrófono, joystick, consola de juegos, antena parabólica, escáner, impresora, dispositivos inalámbricos, etc.

Las realizaciones presentadas en la presente memoria han sido descritas con relación a realizaciones particulares que está previsto, en todos sus aspectos, que sean ilustrativas en vez de restrictivas. Realizaciones alternativas resultarán evidentes para los expertos en la materia a la que pertenece la presente invención sin apartarse de su alcance.

A partir de cuanto antecede, se puede apreciar que esta descripción está bien adaptada para lograr todos los fines y objetos mencionados con anterioridad, definidos junto con otras ventajas que son obvias y que son inherentes a la estructura.

Se comprenderá que determinadas características y sub-combinaciones son de utilidad y pueden ser empleadas sin referencia a otras características o sub-combinaciones. Esto está contemplado por el, y está dentro del, alcance de las reivindicaciones.

REIVINDICACIONES

1. Un sistema de procesamiento de sonido, que comprende:

un primer dispositivo (106) de captura de audio y un segundo dispositivo (108) de captura de audio,

en donde el primer dispositivo (106) de captura de audio está situado en relación de proximidad más cercana a un punto de interés (110) que el segundo dispositivo (108) de captura de audio;

un módulo (114) de detección de actividad de voz para:

recibir un primer y un segundo suministros de audio capturados respectivamente por el primer (106) y el segundo (108) dispositivos de captura de audio;

atenuar al menos una porción del primer suministro de audio en base a una porción correspondiente del segundo suministro de audio para generar un primer suministro de audio atenuado;

atenuar al menos una porción del segundo suministro de audio en base a una porción correspondiente del primer suministro de audio para generar un segundo suministro de audio atenuado;

comparar bandas de frecuencia del primer suministro de audio atenuado con bandas de frecuencia correspondientes del segundo suministro de audio atenuado, y

determinar un nivel de confianza de fuente basado en un número de las bandas de frecuencia a partir del primer suministro de audio atenuado que exceda de un umbral de diferencia predefinido respecto a las bandas de frecuencia correspondientes del segundo suministro de audio atenuado, en donde el nivel de confianza de fuente es indicativo de si el sonido se está originando desde el punto de interés (110).

2. El sistema de procesamiento de sonido de la reivindicación 1, en donde un valor más alto en el nivel de confianza de fuente es más indicativo de sonido dentro del primer suministro de audio atenuado que se origina desde el punto de interés (110) que un nivel más bajo para el nivel de confianza de fuente.

3. El sistema de procesamiento de sonido de la reivindicación 1, en donde, atenuar al menos la porción del primer suministro de audio en base a la porción correspondiente del segundo suministro de audio consiste en atenuar una o más frecuencias contenidas dentro del primer suministro de audio que estén contenidas dentro del segundo suministro de audio, y en donde atenuar al menos la porción del segundo suministro de audio en base a la porción correspondiente del primer suministro de audio consiste en atenuar una o más frecuencias contenidas dentro del segundo suministro de audio que estén contenidas dentro del primer suministro de audio.

4. El sistema de procesamiento de sonido de la reivindicación 1, en donde el módulo (114) de detección de actividad de voz está además para:

sincronizar en el tiempo el primer suministro de audio con el segundo suministro de audio con anterioridad a atenuar al menos la porción del primer suministro de audio, y

sincronizar en el tiempo el segundo suministro de audio con el primer suministro de audio con anterioridad a atenuar al menos la porción del segundo suministro de audio.

5. El sistema de procesamiento de sonido de la reivindicación 1, que comprende además:

un módulo (118) de reconocimiento de voz para:

recibir el primer suministro de audio atenuado;

monitorizar el primer suministro de audio atenuado para identificar uno o más activadores contenidos dentro del primer suministro de audio atenuado, y

provocar que ocurran una o más acciones en respuesta a la identificación de los uno o más activadores.

6. El sistema de procesamiento de sonido de la reivindicación 5, en donde el módulo (114) de detección de actividad de voz está previsto además para: presentar a la salida el primer suministro de audio atenuado para el módulo (118) de reconocimiento de voz en respuesta a una determinación de que el nivel de confianza de fuente excede un límite preconfigurado.

7. El sistema de procesamiento de sonido de la reivindicación 6, en donde el límite preconfigurado varía en base a un nivel de potencia de un dispositivo de computación que alberga el sistema de procesamiento de sonido.

8. El sistema de procesamiento de sonido de la reivindicación 1, en donde el módulo (114) de detección de actividad de voz está previsto además para:

determinar un nivel de confianza de ruido en base a un número de las bandas de frecuencia a partir del primer suministro de audio que estén dentro de un umbral de diferencia predefinido con respecto a las bandas de frecuencia correspondientes del segundo suministro de audio, en donde un valor más alto para el nivel de confianza de ruido es más indicativo de sonido dentro del primer suministro de audio que es ruido, que un valor más bajo del nivel de confianza de ruido.

9. Uno o más medios de almacenaje en ordenador que tienen instrucciones ejecutables con ordenador materializadas en los mismos que, cuando se ejecutan, por medio de uno o más procesadores de un dispositivo de computación, provocan que los uno o más procesadores: lleven a cabo un método para procesar sonido, comprendiendo el método:

filtrar un primer suministro de audio utilizando un segundo suministro de audio para producir un primer suministro de audio filtrado, en donde el primer suministro de audio es capturado por un primer micrófono y el segundo suministro de audio es capturado por un segundo micrófono, estando el primer micrófono en relación de proximidad más cercana a un punto de interés que el segundo micrófono;

filtrar el segundo suministro de audio utilizando el primer suministro de audio para producir un segundo suministro de audio filtrado;

comparar bandas de frecuencia del primer suministro de audio filtrado con bandas de frecuencia correspondientes del segundo suministro de audio filtrado, y

determinar un nivel de confianza de fuente en base a un número de las bandas de frecuencia del primer suministro de audio filtrado que exceda de un umbral de diferencia predefinido con respecto a las bandas de frecuencia correspondientes del segundo suministro de audio filtrado, en donde el nivel de confianza de fuente es indicativo de si el sonido se está originando desde el punto de interés (110).

10. Los uno o más medios de almacenaje en ordenador de la reivindicación 9, comprendiendo además el método enviar el primer suministro de audio filtrado hasta un motor de reconocimiento de voz del dispositivo de computación en respuesta a que el nivel de confianza de fuente exceda de un límite preconfigurado, en donde el límite preconfigurado varía en base a un nivel de potencia del dispositivo de computación.

11. Un método implementado en ordenador para la detección de actividad de voz, que comprende:

recibir un primer suministro de audio capturado por un primer micrófono de un dispositivo de computación, y un segundo suministro de audio capturado por un segundo micrófono del dispositivo de computación, en donde el primer micrófono está en relación de proximidad más cercana a una fuente de interés que el segundo micrófono;

filtrar frecuencias del primer suministro de audio en base a frecuencias correspondientes del segundo suministro de audio para producir un primer suministro de audio filtrado;

filtrar frecuencias del segundo suministro de audio en base a frecuencias correspondientes del primer suministro de audio para producir un segundo suministro de audio filtrado;

comparar bandas de frecuencia del primer suministro de audio filtrado con bandas de frecuencia correspondientes del segundo suministro de audio filtrado, y

determinar un nivel de confianza de fuente en base a un número de las bandas del primer suministro de audio filtrado que exceda de un umbral predefinido de diferencia respecto a las bandas de frecuencia correspondientes del segundo suministro de audio filtrado, en donde un valor más alto del nivel de confianza de fuente es más indicativo de sonido dentro del primer suministro de audio que se origina desde la dirección de la fuente de interés que un valor más bajo del nivel de confianza de fuente.

12. El método implementado en ordenador de la reivindicación 11, en donde la fuente de interés es un usuario del dispositivo de computación, comprendiendo además el método:

enviar el primer suministro de audio filtrado hasta un módulo (118) de reconocimiento de voz del dispositivo de computación en respuesta a una determinación de que el valor para el nivel de confianza de fuente excede un límite preconfigurado, en donde el límite preconfigurado está basado en un nivel de potencia actual del dispositivo de computación, y en donde un límite preconfigurado más alto reduce la cantidad del primer suministro de audio que se presenta a la salida para el módulo (118) de reconocimiento de voz.

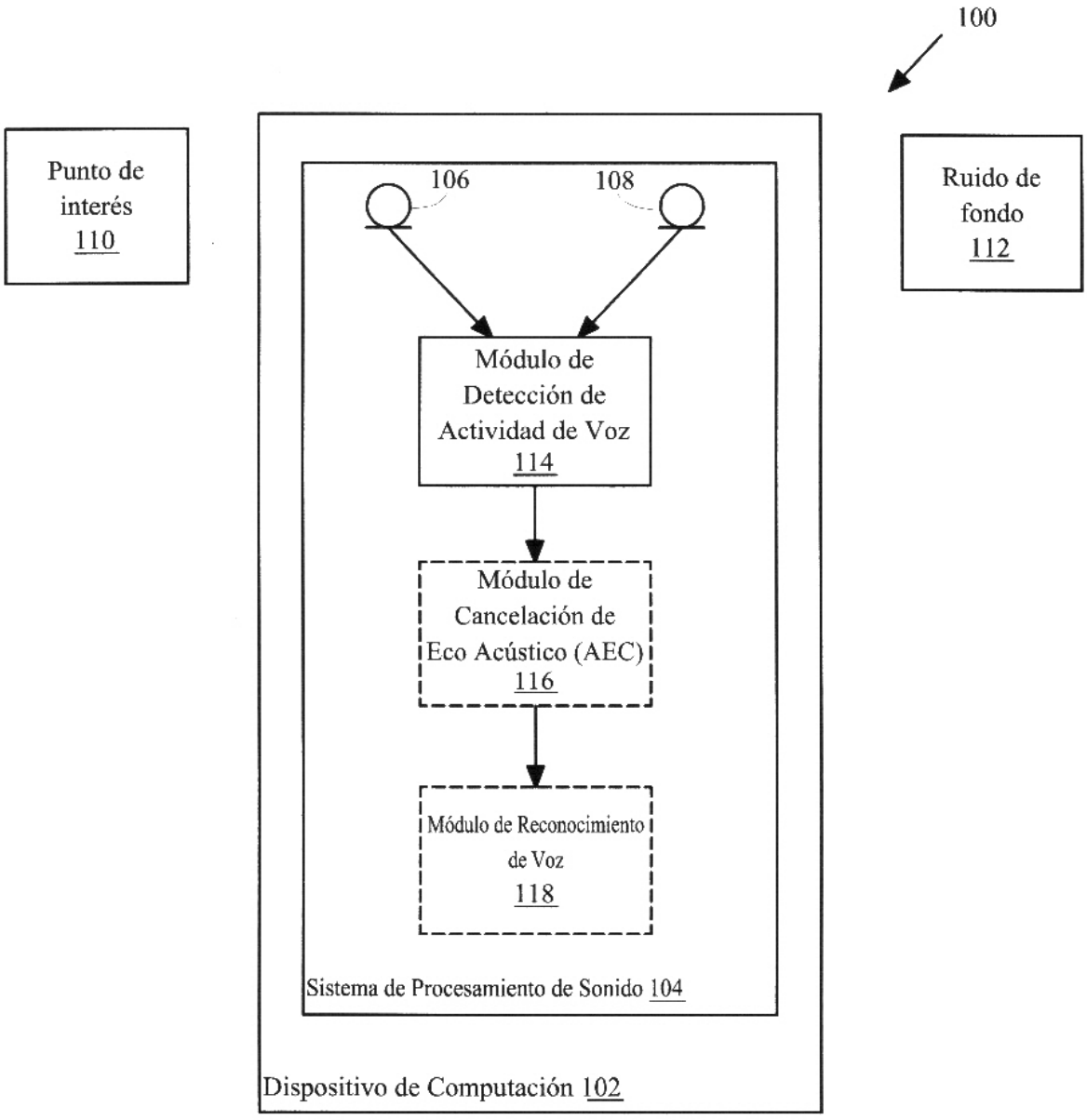
100
Punto de interés 110
Ruido de fondo 112
106
108
Módulo de Detección de Actividad de Voz 114
Módulo de Cancelación de Eco Acústico (AEC) 116
Módulo de Reconocimiento de Voz 118
Sistema de Procesamiento de Sonido 104
Dispositivo de Computación 102

FIG. 1

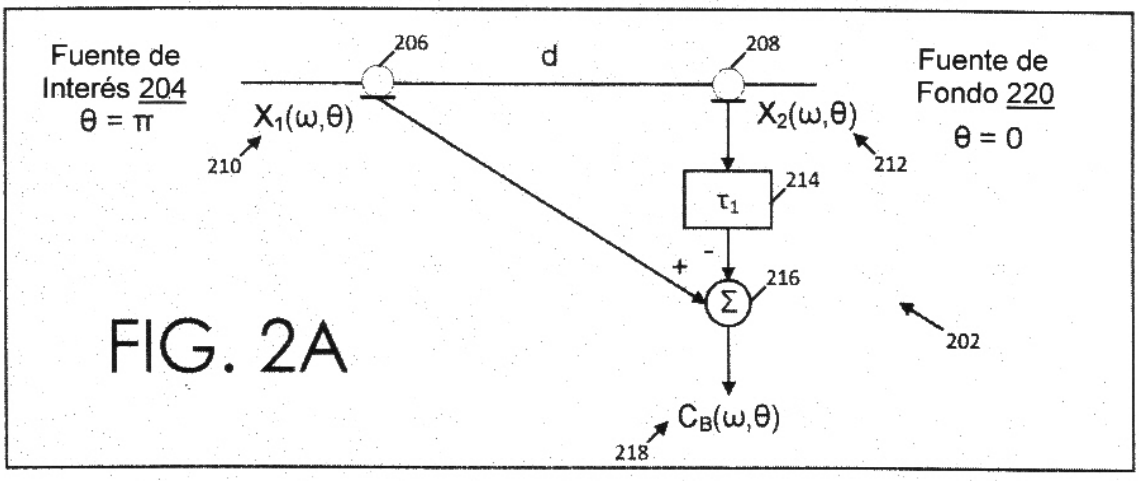

FIG. 2A

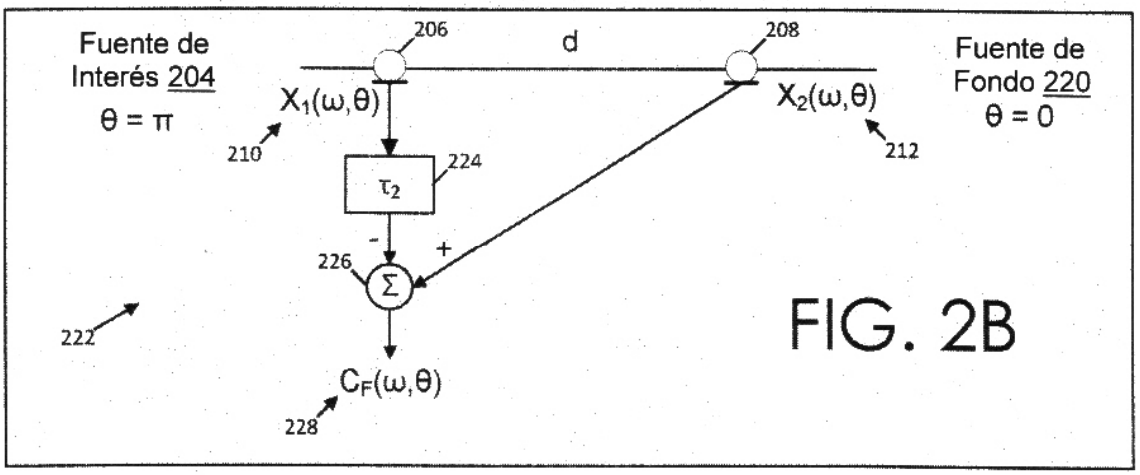

FIG. 2B

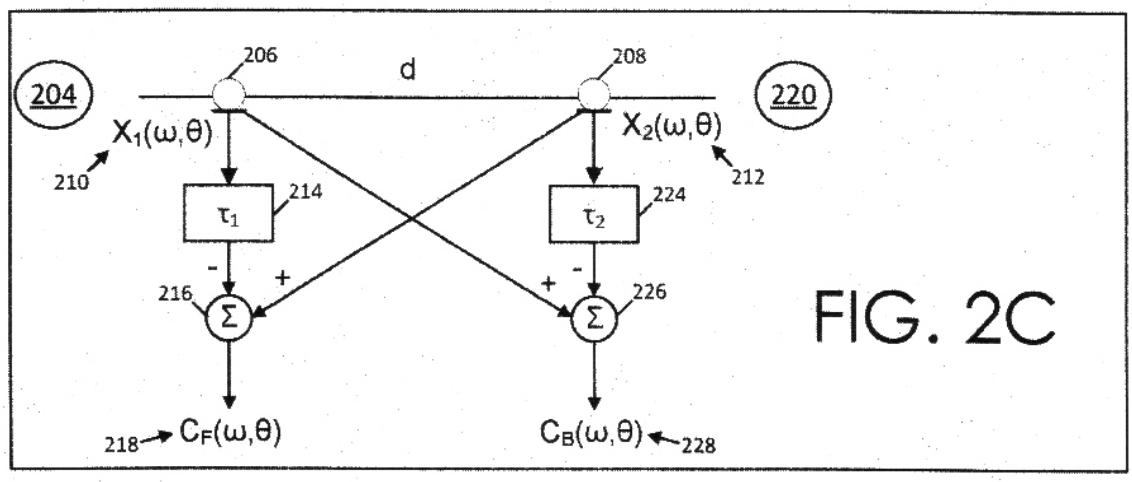

FIG. 2C

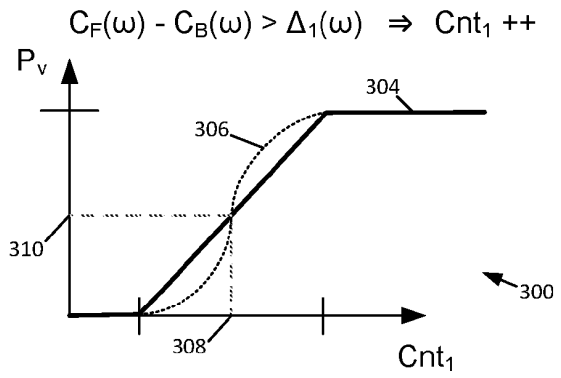

FIG. 3A

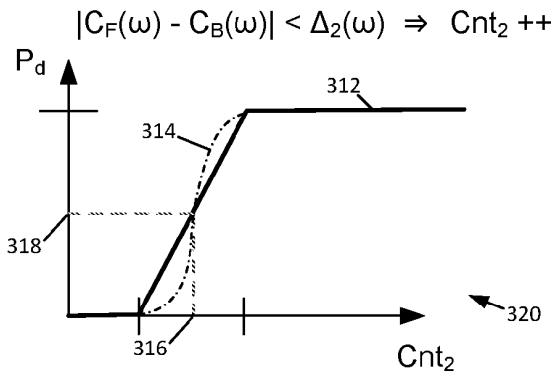

FIG. 3B

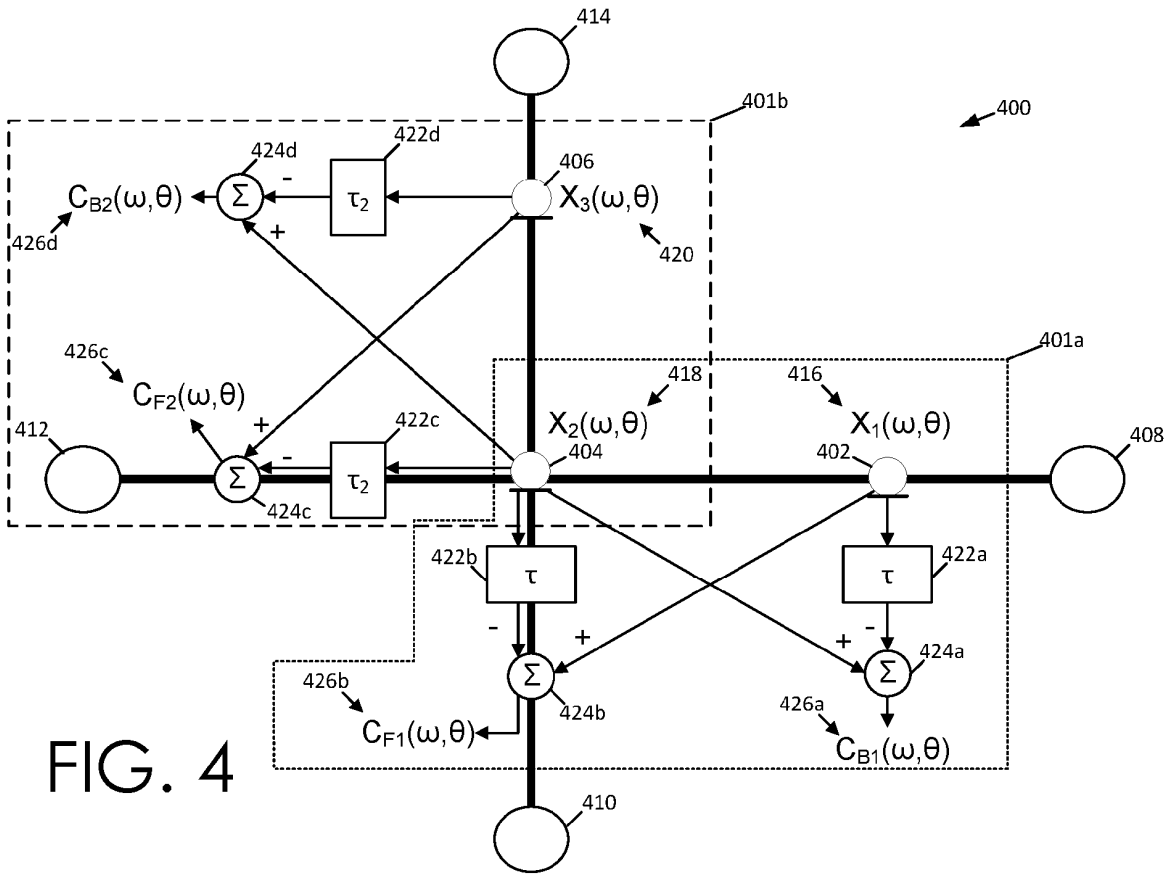

FIG. 4

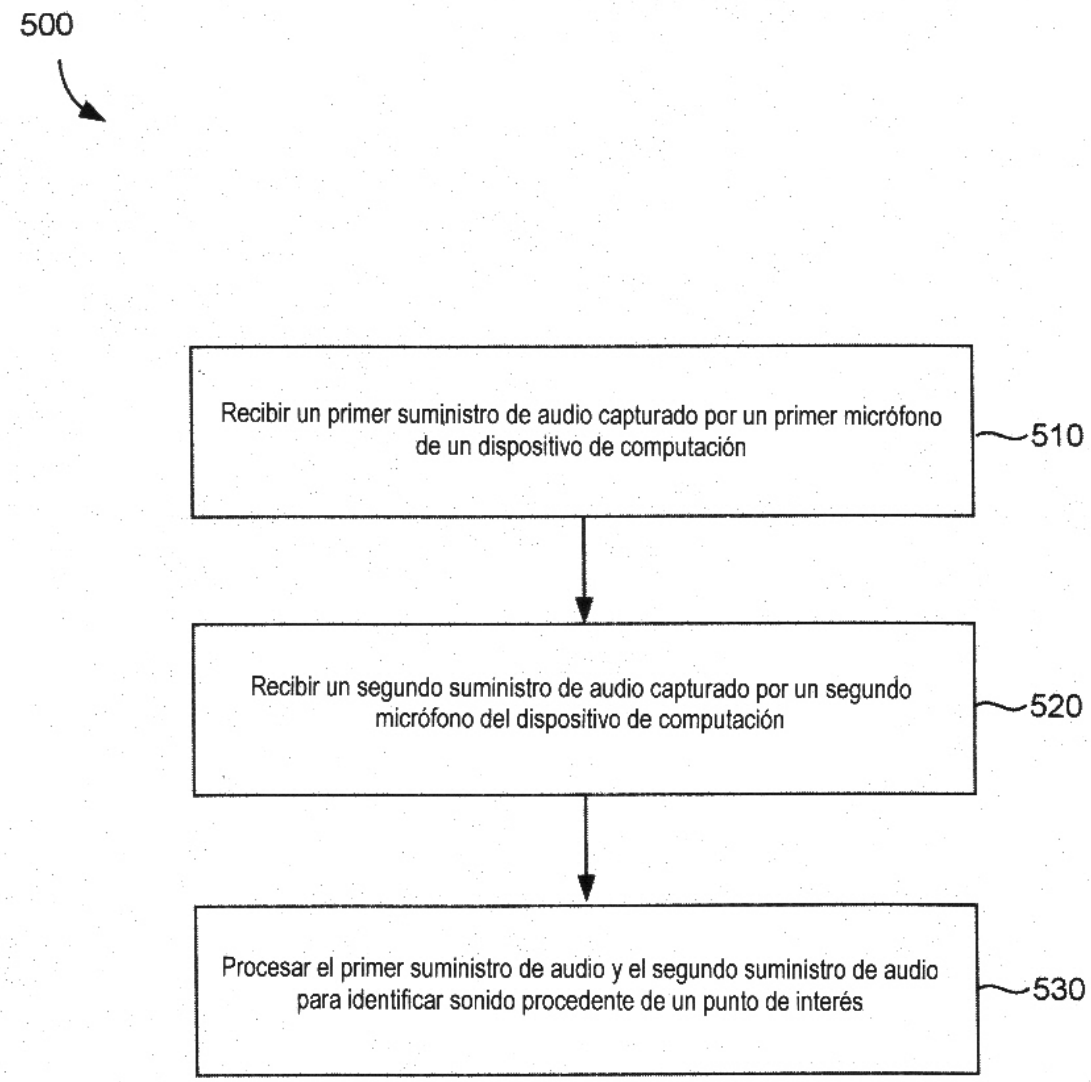
500
Recibir un primer suministro de audio capturado por un primer micrófono de un dispositivo de computación
510
Recibir un segundo suministro de audio capturado por un segundo micrófono del dispositivo de computación
520
Procesar el primer suministro de audio y el segundo suministro de audio para identificar sonido procedente de un punto de interés
530

FIG. 5

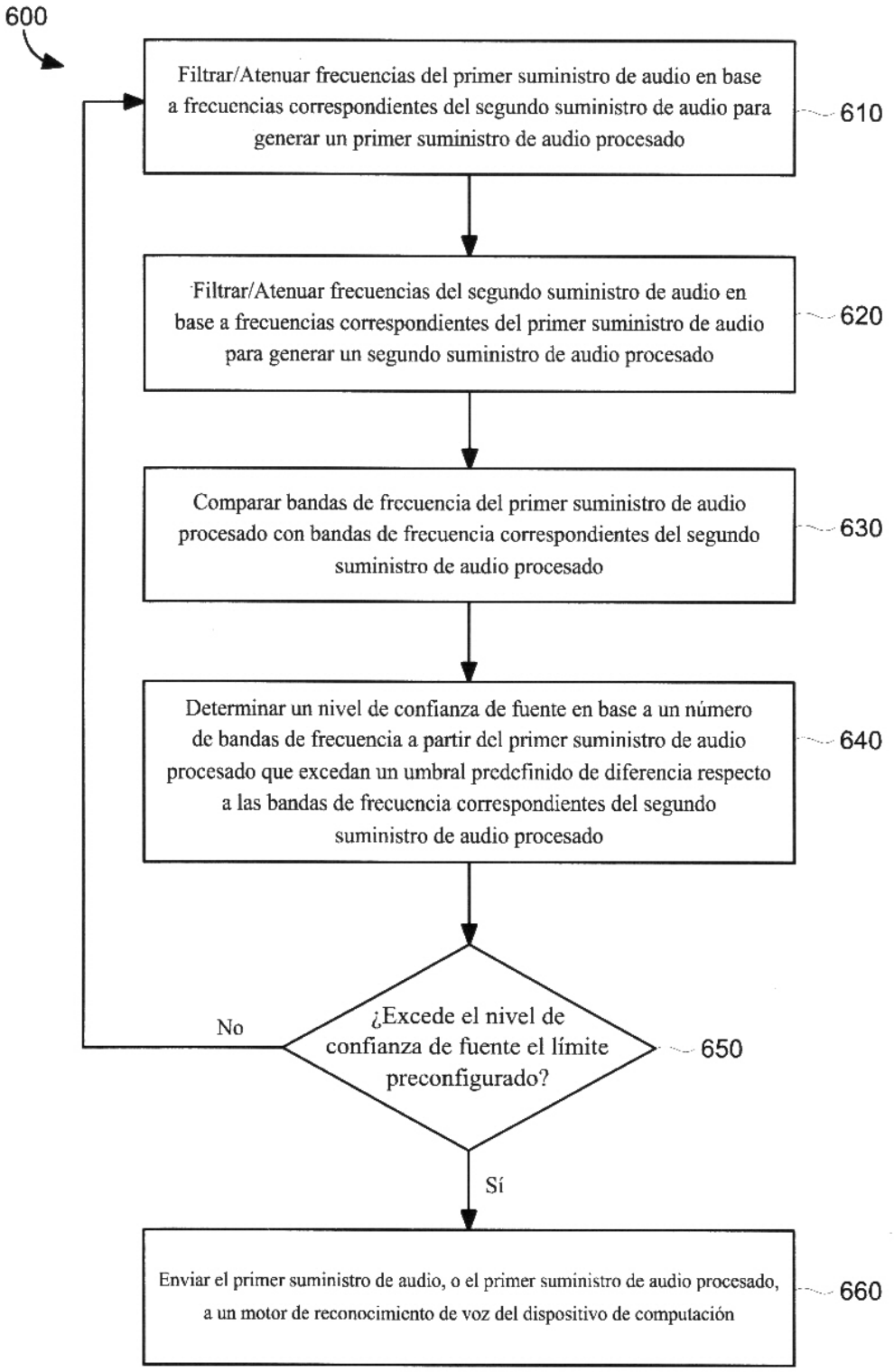
600
Filtrar/Atenuar frecuencias del primer suministro de audio en base a frecuencias correspondientes del segundo suministro de audio para generar un primer suministro de audio procesado
610
Filtrar/Atenuar frecuencias del segundo suministro de audio en base a frecuencias correspondientes del primer suministro de audio para generar un segundo suministro de audio procesado
620
Comparar bandas de frecuencia del primer suministro de audio procesado con bandas de frecuencia correspondientes del segundo suministro de audio procesado
630
Determinar un nivel de confianza de fuente en base a un número de bandas de frecuencia a partir del primer suministro de audio procesado que excedan un umbral predefinido de diferencia respecto a las bandas de frecuencia correspondientes del segundo suministro de audio procesado
640
¿Excede el nivel de confianza de fuente el límite preconfigurado?
650
No
Sí
Enviar el primer suministro de audio, o el primer suministro de audio procesado, a un motor de reconocimiento de voz del dispositivo de computación
660

FIG. 6

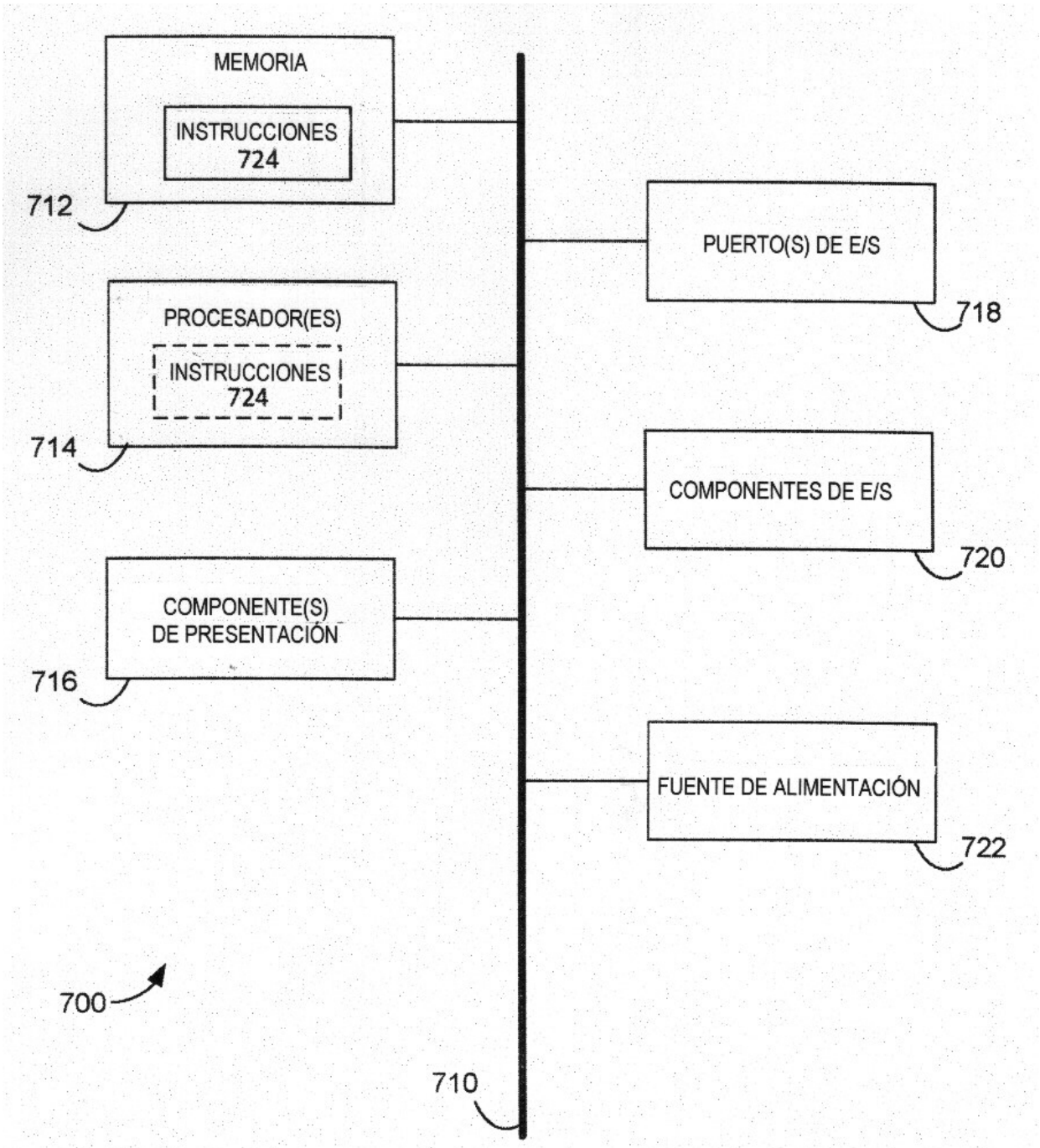
MEMORIA
INSTRUCCIONES 724
712
PROCESADOR(ES)
INSTRUCCIONES 724
714
COMPONENTE(S) DE PRESENTACIÓN
716
PUERTO(S) DE E/S
718
COMPONENTES DE E/S
720
FUENTE DE ALIMENTACIÓN
722
700
710

FIG. 7