(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 775 799**

(51) Int. CI.:

H04M 1/60 (2006.01)
H04R 3/00 (2006.01)
G10L 21/0208 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **14.02.2005 E 05101071 (8)**

(97) Fecha y número de publicación de la concesión europea: **13.11.2019 EP 1569422**

(54) Título: **Método y aparato para la mejora multisensorial del habla en un dispositivo móvil**

(30) Prioridad:

24.02.2004 US 785768

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
28.07.2020

(73) Titular/es:

ZHIGU HOLDINGS LIMITED (100.0%)
Harneys Services (Cayman) Limited, 4th Floor, Harbour Place, 103 South Church Street, George Town, P.O. Box 10240
Grand Cayman KY1-1002, KY

(72) Inventor/es:

SINCLAIR, MICHAEL J.;
HUANG, XUEDONG DAVID y
ZHANG, ZHENGYOU

(74) Agente/Representante:

LINAGE GONZÁLEZ, Rafael

DESCRIPCIÓN

Método y aparato para la mejora multisensorial del habla en un dispositivo móvil

Antecedentes de la invención

La presente invención se refiere a la reducción del ruido. En particular, la presente invención se refiere a eliminar el ruido de las señales del habla recibidas por dispositivos móviles de mano.

Los dispositivos móviles de mano, como los teléfonos móviles y los asistentes personales digitales que proporcionan funciones telefónicas o aceptan la entrada del habla, a menudo se usan en entornos con ruido adverso, como calles concurridas, restaurantes, aeropuertos y automóviles. El fuerte ruido ambiental en estos entornos puede tapar el habla del usuario y dificultar el entendimiento de lo que la persona dice.

Si bien se han desarrollado sistemas de filtrado de ruido que intentan eliminar el ruido en base a un modelo del ruido, estos sistemas no han podido eliminar todo el ruido. En particular, a muchos de estos sistemas les ha resultado difícil eliminar el ruido que consiste en que otras personas hablen en segundo plano. Una razón para esto es que es extremadamente difícil, si no imposible, que estos sistemas determinen que una señal del habla recibida por un micrófono proviene de alguien que no sea la persona que usa el dispositivo móvil.

En cuanto a los auriculares para teléfonos, que se mantienen en posición en la cabeza del usuario al colocar el auricular sobre la cabeza o en el oído del usuario, se han desarrollado sistemas que proporcionan un filtro de ruido más robusto que dependen de tipos de sensores adicionales en el auricular. En un ejemplo, se coloca un sensor de conducción ósea en un extremo del auricular y se presiona y se pone en contacto con la piel que cubre el cráneo, el oído o la mandíbula del usuario mediante la resistencia del auricular. El sensor de conducción ósea detecta vibraciones en el cráneo, el oído o la mandíbula que se crean cuando el usuario habla. Mediante la señal del sensor de conducción ósea, este sistema puede identificar mejor cuando el usuario está hablando y, como resultado, puede filtrar mejor el ruido en la señal del habla.

Aunque estos sistemas funcionan bien para auriculares, donde el contacto entre el sensor de conducción ósea y el usuario se mantiene mediante el diseño mecánico de los auriculares, estos sistemas no pueden usarse directamente en dispositivos móviles de mano porque es difícil para los usuarios mantener el sensor de conducción ósea en la posición correcta y estos sistemas no tienen en cuenta que el sensor de conducción ósea puede no mantenerse en la posición correcta.

El documento WO 2004/012477 A2 describe una interfaz de audio ponible que incluye un soporte para colocar la pluralidad de altavoces yuxtapuestos con y separados de los oídos del usuario.

El documento US 2003/0125081 A1 describe un dispositivo electrónico personal de mano con un primer cuerpo y un segundo cuerpo. El primer cuerpo tiene una pantalla y el segundo cuerpo tiene una pluralidad de entradas manuales.

YANLI ZHENG Y COL.: "Air- and bone-conductive integrated microphones for robust speech detection and enhancement" AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING, 2003. ASRU '03. 2003 IEEE WORKSHOP ON ST. THOMAS, VI, USA NOV. 30-DIC. 3, 2003, PISCATAWAY, NJ, EE. UU. IEEE, 30 de noviembre de 2003 (2003-11-30), páginas 249-254, en el documento XP010713318 describe un dispositivo de hardware que combina un micrófono normal con un micrófono conductor óseo.

El documento JP 2000250577 A describe un dispositivo de reconocimiento de voz, un método de aprendizaje y un dispositivo de aprendizaje que se utilizará en el mismo dispositivo.

El documento JPH 09284877 A describe un sistema de micrófono para obtener un sonido de alta calidad sin verse afectado por el ruido ambiental.

Compendio de la invención

Se proporciona un dispositivo móvil que incluye una entrada de dígitos que puede manipularse con los dedos o el pulgar de un usuario, un micrófono de conducción de aire y un sensor alternativo que proporciona una señal de sensor alternativo indicativa del habla. En algunas realizaciones, el dispositivo móvil también incluye un sensor de proximidad que proporciona una señal de proximidad indicativa de la distancia desde el dispositivo móvil a un objeto. En algunas realizaciones, la señal del micrófono de conducción de aire, la señal del sensor alternativo y la señal de proximidad se usan para formar una estimación de un valor del habla limpia. En realizaciones adicionales, se produce un sonido a través de un altavoz en el dispositivo móvil basado en la cantidad de ruido en el valor del habla limpia. En otras realizaciones, el sonido producido a través del altavoz se basa en la señal del sensor de proximidad.

Breve descripción de los dibujos

La FIG. 1 es una vista en perspectiva de una realización de la presente invención.

La FIG. 2 muestra el teléfono de la FIG. 1 en posición en el lado izquierdo de la cabeza de un usuario.

La FIG. 3 muestra el teléfono de la FIG. 1 en posición en el lado derecho de la cabeza de un usuario.

La FIG. 4 es un diagrama de bloques de un micrófono de conducción ósea.

La FIG. 5 es una vista en perspectiva de una realización alternativa de la presente invención.

La FIG. 6 es un corte transversal de un micrófono de conducción ósea alternativo en una realización de la presente invención.

La FIG. 7 es un diagrama de bloques de un dispositivo móvil en una realización de la presente invención.

La FIG. 8 es un diagrama de bloques de un sistema de procesamiento del habla general de la presente invención.

La FIG. 9 es un diagrama de bloques de un sistema para entrenar parámetros de reducción del ruido en una realización de la presente invención.

La FIG. 10 es un diagrama de flujo para entrenar parámetros de reducción del ruido con el sistema de la FIG. 9.

La FIG. 11 es un diagrama de bloques de un sistema para identificar una estimación de una señal del habla limpia a partir de una señal del habla de prueba de ruido en una realización de la presente invención.

La FIG. 12 es un diagrama de flujo de un método para identificar una estimación de una señal del habla limpia con el sistema de la FIG. 11.

La FIG. 13 es un diagrama de bloques de un sistema alternativo para identificar una estimación de una señal del habla limpia.

La FIG. 14 es un diagrama de bloques de un segundo sistema alternativo para identificar una estimación de una señal del habla limpia.

La FIG. 15 es un diagrama de flujo de un método para identificar una estimación de una señal del habla limpia con el sistema de la FIG. 14.

La FIG. 16 es una vista en perspectiva de una realización adicional de un dispositivo móvil de la presente invención.

Descripción detallada de realizaciones ilustrativas

Las realizaciones de la presente invención proporcionan dispositivos móviles de mano que contienen tanto un micrófono de conducción de aire como un sensor alternativo que puede usarse en la detección del habla y el filtrado de ruido. La FIG. 1 proporciona una realización de ejemplo en la que el dispositivo móvil de mano es un teléfono móvil 100. El teléfono móvil 100 incluye un teclado 102, una pantalla 104, un control del cursor 106, un micrófono de conducción de aire 108, un altavoz 110, dos micrófonos de conducción ósea 112 y 114 y, opcionalmente, un sensor de proximidad 116.

El panel táctil 102 permite que el usuario introduzca números y letras en el teléfono móvil. En otras realizaciones, el panel táctil 102 se combina con la pantalla 104 en forma de una pantalla táctil. El control del cursor 106 permite al usuario resaltar y seleccionar información en la pantalla 104 y desplazarse a través de imágenes y páginas que son más grandes que la pantalla 104.

Como se muestra en las FIGS. 2 y 3, cuando el teléfono móvil 100 se coloca en la posición estándar para conversar por teléfono, el altavoz 110 se coloca cerca del oído izquierdo 200 o el oído derecho 300 del usuario, y el micrófono de conducción de aire 108 se coloca cerca de la boca 202 del usuario. Cuando el teléfono se coloca cerca del oído izquierdo del usuario, como en la FIG. 2, el micrófono de conducción ósea 114 entra en contacto con el cráneo o el oído del usuario y produce una señal de sensor alternativo que puede usarse para eliminar el ruido de la señal del habla recibida por el micrófono de conducción de aire 108. Cuando el teléfono se coloca cerca del oído derecho del usuario, como en la FIG. 3, el micrófono de conducción ósea 112 entra en contacto con el cráneo o el oído del usuario y produce una señal de sensor alternativo que puede usarse para eliminar el ruido de la señal del habla.

El sensor de proximidad opcional 116 indica qué tan cerca está el teléfono del usuario. Como se analiza más adelante, esta información se utiliza para ponderar la contribución de los micrófonos de conducción ósea en la producción del valor del habla limpia. En general, si el detector de proximidad detecta que el teléfono está al lado del usuario, las señales del micrófono de conducción ósea se ponderan más que si el teléfono está a cierta distancia del usuario. Este ajuste refleja el hecho de que la señal del micrófono de conducción ósea es más indicativa de que el usuario habla cuando está en contacto con el usuario. Cuando está apartado del usuario, es más susceptible al ruido ambiental. El sensor de proximidad se usa en realizaciones de la presente invención porque los usuarios no siempre sostienen el teléfono presionado contra sus cabezas.

La FIG. 4 muestra una realización de un sensor de conducción ósea 400 de la presente invención. En el sensor 400,

un puente de elastómero blando 402 está adherido a un diafragma 404 de un micrófono de conducción de aire normal 406. Este puente blando 402 conduce vibraciones desde el contacto con la piel 408 del usuario directamente al diafragma 404 del micrófono 406. El movimiento del diafragma 404 se convierte en una señal eléctrica mediante un transductor 410 en el micrófono 406.

La FIG. 5 proporciona una realización alternativa de teléfono móvil 500 del dispositivo móvil de mano de la presente invención. El teléfono móvil 500 incluye un teclado 502, una pantalla 504, un control del cursor 506, un micrófono de conducción de aire 508, un altavoz 510 y una combinación del micrófono de conducción ósea y sensor de proximidad 512.

Como se muestra en el corte transversal de la FIG. 6, la combinación del micrófono de conducción ósea y el sensor de proximidad 512 consiste en una almohadilla suave, de relleno de medio (con fluido o elastómero) 600 que tiene una superficie exterior 602 diseñada para entrar en contacto con al usuario cuando el usuario coloca el teléfono contra su oído. La almohadilla 600 forma un anillo alrededor de una abertura que proporciona una vía de paso para el sonido del altavoz 510, que se encuentra en la abertura o directamente debajo de la abertura dentro del teléfono 500. La almohadilla 600 no está limitada a esta forma y se puede usar cualquier forma para la almohadilla. En general, sin embargo, se prefiere si la almohadilla 600 incluye porciones a la izquierda y derecha del altavoz 510, de modo que al menos una parte de la almohadilla 600 esté en contacto con el usuario, independientemente del oído en el que el usuario coloque el teléfono. Las porciones de la almohadilla pueden ser externamente continuas o pueden estar externamente separadas pero conectadas de manera fluida entre sí dentro del teléfono.

Un transductor electrónico de presión 604 está conectado hidráulicamente al fluido o elastómero en la almohadilla 600 y convierte la presión del fluido en la almohadilla 600 en una señal eléctrica en el conductor 606. Los ejemplos de transductor electrónico de presión 604 incluyen transductores basados en tecnología MEMS. En general, el transductor de presión 604 debería tener una respuesta de alta frecuencia.

La señal eléctrica en el conductor 606 incluye dos componentes, un componente de CC y un componente de CA. El componente de CC proporciona una señal del sensor de proximidad porque la presión estática dentro de la almohadilla 600 será mayor cuando el teléfono se presiona contra el oído del usuario que cuando el teléfono está a cierta distancia del oído del usuario. El componente de CA de la señal eléctrica proporciona una señal de micrófono de conducción ósea porque las vibraciones en los huesos del cráneo, la mandíbula o el oído del usuario crean fluctuaciones de presión en la almohadilla 600 que se convierten en una señal eléctrica de CA mediante el transductor de presión 604. En una realización, se aplica un filtro a la señal eléctrica para permitir que pase el componente de CC de la señal y los componentes de CA por encima de una frecuencia mínima.

Aunque se han descrito anteriormente dos ejemplos de sensores de conducción ósea, otras formas para el sensor de conducción ósea se encuentran dentro del alcance de la presente invención.

La FIG. 7 es un diagrama de bloques de un dispositivo móvil 700, en una realización de la presente invención. El dispositivo móvil 700 incluye un microprocesador 702, memoria 704, interfaz de entrada/salida (E/S) 706 y una interfaz de comunicación 708 para comunicarse con ordenadores remotos, redes de comunicación u otros dispositivos móviles. En una realización, los componentes mencionados anteriormente se acoplan para comunicarse entre sí a través de un bus 710 adecuado.

La memoria 704 puede implementarse como memoria electrónica no volátil, tal como memoria de acceso aleatorio (RAM) con un módulo de reserva de batería (no se muestra) de manera que la información almacenada en la memoria 704 no se pierda cuando la alimentación general del dispositivo móvil 700 se desconecte. De forma alternativa, toda o porciones de la memoria 704 pueden ser memorias extraíbles volátiles o no volátiles. Una porción de memoria 704 se asigna preferiblemente como memoria direccionable para la ejecución del programa, mientras que otra porción de memoria 704 se usa preferiblemente para almacenamiento, por ejemplo, simular almacenamiento en una unidad de disco.

La memoria 704 incluye un sistema operativo 712, programas de aplicación 714, así como un almacén de objetos 716. En funcionamiento, el sistema operativo 712 es ejecutado preferiblemente por el procesador 702 desde la memoria 704. El sistema operativo 712, en una realización preferida, es un sistema operativo de la marca WINDOWS® CE disponible comercialmente de Microsoft Corporation. El sistema operativo 712 está diseñado preferiblemente para dispositivos móviles e implementa características de base de datos que pueden ser utilizadas por las aplicaciones 714 a través de un conjunto de interfaces y métodos de programación de aplicaciones expuestos. Los objetos en el almacén de objetos 716 son mantenidos por las aplicaciones 714 y el sistema operativo 712, al menos parcialmente en respuesta a las llamadas a las interfaces y métodos de programación de aplicaciones expuestos.

La interfaz de comunicación 708 representa numerosos dispositivos y tecnologías que permiten que el dispositivo móvil 700 envíe y reciba información. En realizaciones de teléfonos móviles, la interfaz de comunicación 708 representa una interfaz de red de teléfono celular que interactúa con una red de teléfono celular para permitir que se realicen y se reciban llamadas. Otros dispositivos posiblemente representados por la interfaz de comunicación 708 incluyen módems cableados e inalámbricos, receptores de satélite y sintonizadores de radiodifusión, por nombrar

algunos. El dispositivo móvil 700 también se puede conectar directamente a un ordenador para intercambiar datos con el mismo. En dichos casos, la interfaz de comunicación 708 puede ser un transceptor de infrarrojos o una conexión de comunicación en serie o en paralelo, todos los cuales son capaces de transmitir información de emisión en continuo.

Las instrucciones ejecutables por ordenador que son ejecutadas por el procesador 702 para implementar la presente invención pueden almacenarse en la memoria 704 o recibirse a través de la interfaz de comunicación 708. Estas instrucciones se encuentran en un medio legible por ordenador, que, sin limitación, puede incluir medios de almacenamiento informático y medios de comunicación.

Los medios de almacenamiento informáticos incluyen medios volátiles y no volátiles, extraíbles y no extraíbles implementados en cualquier método o tecnología para el almacenamiento de información, tales como instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos. Los medios de almacenamiento informático incluyen, entre otros, RAM, ROM, EEPROM, memoria flash u otra tecnología de memoria, CD-ROM, discos versátiles digitales (DVD) u otro almacenamiento en disco óptico, casetes magnéticos, cinta magnética, almacenamiento en disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio que pueda usarse para almacenar la información deseada y al que se pueda acceder.

Los medios de comunicación típicamente incorporan instrucciones legibles por 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 entrega de información. El término "señal de datos modulada" significa una señal que tiene una o más de sus características establecida o cambiada de manera que codifique información en la señal. A modo de ejemplo, y sin limitación, los medios de comunicación incluyen medios cableados como una red cableada o conexión cableada directa, y medios inalámbricos como medios acústicos, RF, infrarrojos y otros medios inalámbricos. Las combinaciones de cualquiera de lo anterior deberían estar incluidos también dentro del alcance de los medios legibles por ordenador.

La interfaz de entrada/salida 706 representa interfaces para una colección de dispositivos de entrada y salida, incluido el altavoz 730, la entrada de dígitos 732 (como uno o un conjunto de botones, una pantalla táctil, una bola de seguimiento, una alfombrilla de ratón, un rodillo o una combinación de estos componentes que pueden manipularse con el pulgar o el dedo del usuario), pantalla 734, micrófono de conducción de aire 736, sensor alternativo 738, sensor alternativo 740 y sensor de proximidad 742. En una realización, los sensores alternativos 738 y 740 son micrófonos de conducción ósea. Los dispositivos enumerados anteriormente son a modo de ejemplo y no todos necesitan estar presentes en el dispositivo móvil 700. Además, en al menos una realización, el sensor alternativo y el sensor de proximidad se combinan como un único sensor que proporciona una señal del sensor de proximidad y una señal de sensor alternativo. Estas señales pueden colocarse en líneas de conducción separadas o pueden ser componentes de una señal en una única línea. Además, otros dispositivos de entrada/salida se pueden conectar a, o encontrar con, el dispositivo móvil 700 dentro del alcance de la presente invención.

La FIG. 8 proporciona un diagrama de bloques básico de un sistema de procesamiento del habla de realizaciones de la presente invención. En la FIG. 8, un altavoz 800 genera una señal del habla 802 que es detectada por un micrófono de conducción de aire 804 y uno o ambos de un sensor alternativo 806 y un sensor alternativo 807. Un ejemplo de un sensor alternativo es un sensor de conducción ósea que se encuentra en, o adyacente, a un hueso facial o del cráneo del usuario (como el hueso de la mandíbula) o en el oído del usuario y que detecta las vibraciones del oído, el cráneo o la mandíbula que corresponden al habla generado por el usuario. Otro ejemplo de un sensor alternativo es un sensor infrarrojo que apunta hacia y detecta el movimiento de la boca del usuario. Obsérvese que, en algunas realizaciones, solamente estará presente un sensor alternativo. El micrófono de conducción de aire 804 es el tipo de micrófono que se utiliza comúnmente para convertir ondas de radio de audio en señales eléctricas.

El micrófono de conducción de aire 804 también recibe ruido 808 generado por una o más fuentes de ruido 810. Dependiendo del tipo de sensor alternativo y el nivel del ruido, el ruido 808 también puede ser detectado por sensores alternativos 806 y 807. Sin embargo, según las realizaciones de la presente invención, los sensores alternativos 806 y 807 son típicamente menos sensibles al ruido ambiental que el micrófono de conducción de aire 804. Por lo tanto, las señales de sensor alternativo 812 y 813 generadas por los sensores alternativos 806 y 807, respectivamente, en general incluyen menos ruido que la señal del micrófono de conducción de aire 814 generada por el micrófono de conducción de aire 804.

Si hay dos sensores alternativos, como dos sensores de conducción ósea, las señales del sensor 812 y 813 se pueden proporcionar opcionalmente a una unidad de comparación/selección 815. La unidad de comparación/selección 815 compara la intensidad de las dos señales y selecciona la señal más fuerte como su salida 817. La señal más débil no se transmite para su procesamiento posterior. En cuanto a realizaciones de teléfonos móviles, tales como el teléfono móvil de las FIGS. 1-3, la unidad de comparación/selección 815 normalmente seleccionará la señal generada por el sensor de conducción ósea que está en contacto con la piel del usuario. Así, en la FIG. 2, se seleccionará la señal del sensor de conducción ósea 114 y en la FIG. 3, se seleccionará la señal del sensor de conducción ósea 112.

La señal de sensor alternativo 817 y la señal del micrófono de conducción de aire 814 se proporcionan a un estimador de la señal limpia 816, que estima una señal del habla limpia 818 a través de un proceso que se comenta a continuación en detalle. Opcionalmente, el estimador de la señal limpia 816 también recibe una señal de proximidad 830 desde un sensor de proximidad 832 que se utiliza para estimar la señal limpia 818. Como se ha indicado anteriormente, el sensor de proximidad puede combinarse con una señal de sensor alternativo en algunas realizaciones. Se proporciona una estimación de la señal limpia 818 a un proceso del habla 820. La señal del habla limpia 818 puede ser una señal en un dominio del tiempo filtrada o un vector en un dominio de características. Si la estimación de la señal limpia 818 es una señal en el dominio del tiempo, el proceso del habla 820 puede adoptar la forma de un oyente, un transmisor de teléfono celular, un sistema de codificación del habla o un sistema de reconocimiento del habla. Si la señal del habla limpia 818 es un vector en un dominio de características, el proceso del habla 820 será típicamente un sistema de reconocimiento del habla.

El estimador de la señal limpia 816 también produce una estimación del ruido 819, que indica el ruido estimado que está en la señal del habla limpia 818. La estimación del ruido 819 se proporciona a un generador de tono lateral 821, que genera un tono a través de los altavoces del dispositivo móvil en base a la estimación del ruido 819. En particular, el generador de tono lateral 821 aumenta el volumen del tono lateral a medida que aumenta la estimación del ruido 819.

El tono lateral proporciona realimentación al usuario que indica si el usuario está sosteniendo el dispositivo móvil en la mejor posición para aprovechar el sensor alternativo. Por ejemplo, si el usuario no está presionando el sensor de conducción ósea contra su cabeza, el estimador de la señal limpia recibirá una señal de sensor alternativo pobre y producirá una señal limpia ruidosa 818 debido a la señal del sensor alternativo pobre. Esto dará como resultado un tono lateral más fuerte. A medida que el usuario pone el sensor de conducción ósea en contacto con su cabeza, la señal del sensor alternativo mejorará, reduciendo así el ruido en la señal limpia 818 y el volumen del tono lateral. Por lo tanto, un usuario puede aprender rápidamente cómo sostener el teléfono para reducir mejor el ruido en la señal limpia en base a la realimentación en el tono lateral.

En realizaciones alternativas, el tono lateral se genera en base a la señal del sensor de proximidad 830 del sensor de proximidad 832. Cuando el sensor de proximidad indica que el teléfono está en contacto o extremadamente cerca de la cabeza del usuario, el volumen del tono lateral será bajo. Cuando el sensor de proximidad indica que el teléfono está lejos de la cabeza del usuario, el tono lateral será más fuerte.

La presente invención utiliza varios métodos y sistemas para estimar el habla limpia utilizando la señal del micrófono de conducción de aire 814, la señal del sensor alternativo 817 y, opcionalmente, la señal del sensor de proximidad 830. Un sistema utiliza datos de entrenamiento estéreo para entrenar vectores de corrección para la señal del sensor alternativo. Cuando estos vectores de corrección se añaden más tarde a un vector de sensor alternativo de prueba, proporcionan una estimación de un vector de señal limpia. Una extensión adicional de este sistema es realizar un seguimiento primero de las distorsiones variables en el tiempo y luego incorporar esta información en el cálculo de los vectores de corrección y en la estimación del habla limpia.

Un segundo sistema proporciona una interpolación entre la estimación de la señal limpia generada por los vectores de corrección y una estimación formada restando una estimación del ruido actual en la señal de prueba de conducción de aire de la señal de conducción de aire. Un tercer sistema utiliza la señal del sensor alternativo para estimar el tono de la señal del habla y luego utiliza el tono estimado para identificar una estimación de la señal del habla limpia. Cada uno de estos sistemas se analiza por separado a continuación.

Entrenamiento de vectores de corrección estéreo

Las FIGS. 9 y 10 proporcionan un diagrama de bloques y un diagrama de flujo para entrenar vectores de corrección estéreo para las dos realizaciones de la presente invención que se basan en vectores de corrección que generan una estimación del habla limpia.

El método de identificación de vectores de corrección comienza en la etapa 1000 de la FIG. 10, donde una señal de micrófono de conducción de aire "limpia" se convierte en una secuencia de vectores de características. Para hacer esto, un altavoz 900 de la FIG. 9, habla en un micrófono de conducción de aire 910, que convierte las ondas de audio en señales eléctricas. Las señales eléctricas son luego muestreadas por un convertidor analógico-digital 914 que genera una secuencia de valores digitales, que son agrupados en tramas de valores por un constructor de tramas 916. En una realización, el convertidor A-D 914 muestrea la señal analógica a 16 kHz y 16 bits por muestra, creando así 32 kilobytes de datos del habla por segundo y el constructor de tramas 916 crea una nueva trama cada 10 milisegundos que incluye 25 milisegundos de datos de valor.

Cada trama de datos proporcionada por el constructor de tramas 916 se convierte en un vector de características mediante un extractor de características 918. En una realización, el extractor de características 918 forma características cepstrales. Ejemplos de dichas características incluyen cepstrum derivado de LPC y coeficientes de cepstrum de la frecuencia de Mel. Los ejemplos de otros posibles módulos de extracción de características que pueden usarse con la presente invención incluyen módulos para realizar la codificación predictiva lineal (LPC), la predicción lineal perceptiva (PLP) y la extracción de características del modelo auditivo. Obsérvese que la invención

no se limita a estos módulos de extracción de características y que se pueden usar otros módulos dentro del contexto de la presente invención.

En la etapa 1002 de la FIG. 10, una señal de sensor alternativo se convierte en vectores de características. Aunque se muestra que la conversión de la etapa 1002 se produce después de la conversión de la etapa 1000, cualquier parte de la conversión puede realizarse antes, durante o después de la etapa 1000 según la presente invención. La conversión de la etapa 1002 se realiza a través de un proceso similar al descrito anteriormente para la etapa 1000.

En la realización de la FIG. 9, este proceso comienza cuando los sensores alternativos 902 y 903 detectan un evento físico asociado con la producción del habla en el altavoz 900, tal como una vibración ósea o un movimiento facial. Debido a que los sensores alternativos 902 y 903 están separados en el dispositivo móvil, no detectarán los mismos valores en relación con la producción del habla. Los sensores alternativos 902 y 903 convierten el evento físico en señales eléctricas analógicas. Estas señales eléctricas se proporcionan a una unidad de comparación/selección 904, que identifica la más fuerte de las dos señales y proporciona la señal más fuerte en su salida. Obsérvese que, en algunas realizaciones, solamente se usa un sensor alternativo. En dichos casos, la unidad de comparación/selección 904 no está presente.

La señal analógica seleccionada es muestreada por un convertidor analógico-digital 905. Las características de muestreo para el convertidor A/D 905 son las mismas que las descritas anteriormente para el convertidor A/D 914. Las muestras proporcionadas por el convertidor A/D 905 son recogidas en tramas por un constructor de tramas 906, que actúa de manera similar al constructor de tramas 916. Las tramas de muestras se convierten luego en vectores de características mediante un extractor de características 908, que utiliza el mismo método de extracción de características que el extractor de características 918.

Los vectores de características para la señal del sensor alternativo y la señal conductora de aire se proporcionan a un entrenador de reducción del ruido 920 en la FIG. 9. En la etapa 1004 de la FIG. 10, el entrenador de reducción del ruido 920 agrupa los vectores de características para la señal del sensor alternativo en componentes de la mezcla. Esta agrupación se puede hacer agrupando vectores de características similares mediante una técnica de entrenamiento de máxima verosimilitud o agrupando vectores de características que representan una sección temporal de la señal del habla. Los expertos en la técnica reconocerán que pueden usarse otras técnicas para agrupar los vectores de características y que las dos técnicas enumeradas anteriormente solamente se proporcionan como ejemplos.

El entrenador de reducción del ruido 920 luego determina un vector de corrección, r_s, para cada componente de la mezcla, s, en la etapa 1008 de la FIG. 10. En una realización, el vector de corrección para cada componente de la mezcla se determina con el criterio de máxima verosimilitud. Mediante esta técnica, el vector de corrección se calcula como:

$$r_s = \frac{\sum_t p(s \mid b_t)(x_t - b_t)}{\sum_t p(s \mid b_t)} \qquad \textbf{EC. 1}$$

donde x_t es el valor del vector de conducción de aire para la trama t y b_t es el valor del vector de sensor alternativo para la trama t. En la ecuación 1:

$$p(s \mid b_t) = \frac{p(b_t \mid s)p(s)}{\sum_s p(b_t \mid s)p(s)} \qquad \textbf{EC. 2}$$

donde p(s) es simplemente uno sobre el número de componentes de la mezcla y $p(b_t|s)$ se modela como una distribución gaussiana:

$$p(b_t \mid s) = N(b_t; \mu_b, \Gamma_b) \qquad \textbf{EC. 3}$$

con la media μ_b y la varianza Γ_b entrenada mediante un algoritmo de maximización de expectativas (EM) donde cada iteración consiste en las siguientes etapas:

$$\gamma_s(t) = p(s \mid b_t) \qquad \textbf{EC. 4}$$

$$\mu_s = \frac{\sum_t \gamma_s(t) b_t}{\sum_t \gamma_s(t)} \qquad \textbf{EC. 5}$$

$$\Gamma_s = \frac{\sum_t \gamma_s(t)(b_t - \mu_s)(b_t - \mu_s)^T}{\sum_t \gamma_s(t)} \qquad \textbf{EC. 6}$$

EC.4 es la etapa E en el algoritmo de EM, que utiliza los parámetros estimados previamente. EC.5 y EC.6 son la etapa M, que actualiza los parámetros utilizando los resultados de la etapa E.

Las etapas E y M del algoritmo iteran hasta que se determinan valores estables para los parámetros del modelo. Estos parámetros se utilizan para evaluar la ecuación 1 y formar los vectores de corrección. Los vectores de corrección y los parámetros del modelo se almacenan luego en un almacén de parámetros de reducción del ruido 922.

Después de que se haya determinado un vector de corrección para cada componente de la mezcla en la etapa 1008, se completa el proceso de entrenamiento del sistema de reducción del ruido de la presente invención. Una vez que se ha determinado un vector de corrección para cada mezcla, los vectores pueden usarse en una técnica de reducción del ruido de la presente invención. A continuación se comentan dos técnicas de reducción del ruido separadas que usan los vectores de corrección.

Reducción de ruido mediante el vector de corrección y la estimación del ruido

Un sistema y método que reduce el ruido en una señal del habla ruidosa en base a vectores de corrección y una estimación del ruido se muestra en el diagrama de bloques de la FIG. 11 y el diagrama de flujo de la FIG. 12, respectivamente.

En la etapa 1200, una señal de prueba de audio detectada por un micrófono de conducción de aire 1104 se convierte en vectores de características. La señal de prueba de audio recibida por el micrófono 1104 incluye el habla de un altavoz 1100 y el ruido añadido procedente de una o más fuentes de ruido 1102. La señal de prueba de audio detectada por el micrófono 1104 se convierte en una señal eléctrica que se proporciona al convertidor analógico-digital 1106.

El convertidor A-D 1106 convierte la señal analógica del micrófono 1104 en una serie de valores digitales. En varias realizaciones, el convertidor A-D 1106 muestrea la señal analógica a 16 kHz y 16 bits por muestra, creando así 32 kilobytes de datos del habla por segundo. Estos valores digitales se proporcionan a un constructor de tramas 1108, que, en una realización, agrupa los valores en tramas de 25 milisegundos que comienzan con una separación de 10 milisegundos.

Las tramas de datos creadas por el constructor de tramas 1108 se proporcionan al extractor de características 1110, que extrae una característica de cada trama. En una realización, este extractor de características es diferente de los extractores de características 908 y 918 que se usaron para entrenar los vectores de corrección. En particular, en esta realización, el extractor de características 1110 produce valores del espectro de potencia en lugar de valores cepstrales. Las características extraídas se proporcionan a un estimador de la señal limpia 1122, una unidad de detección del habla 1126 y un entrenador del modelo de ruido 1124.

En la etapa 1202, un evento físico, tal como una vibración ósea o un movimiento facial, asociado con la producción del habla en el altavoz 1100 se convierte en un vector de características. Aunque se muestra como una etapa separada en la FIG. 12, los expertos en la técnica reconocerán que porciones de esta etapa pueden realizarse al mismo tiempo que la etapa 1200. Durante la etapa 1202, el evento físico es detectado por uno o ambos sensores alternativos 1112 y 1114. Los sensores alternativos 1112 y 1114 generan señales eléctricas analógicas basadas en el evento físico. Las señales analógicas se proporcionan a una unidad de comparación y selección 1115, que selecciona la señal de mayor magnitud como su salida. Obsérvese que, en algunas realizaciones, solamente se proporciona un sensor alternativo. En dichas realizaciones, no se necesita una unidad de comparación y selección 1115.

La señal analógica seleccionada se convierte en una señal digital mediante el convertidor analógico-digital 1116 y las muestras digitales resultantes se agrupan en tramas con el constructor de tramas 1118. En una realización, el convertidor analógico-digital 1116 y el constructor de tramas 1118 funcionan de manera similar al convertidor analógico-digital 1106 y al constructor de tramas 1108.

Las tramas de valores digitales se proporcionan a un extractor de características 1120, que utiliza la misma técnica de extracción de características que se ha utilizado para entrenar los vectores de corrección. Como se ha mencionado anteriormente, los ejemplos de dichos módulos de extracción de características incluyen módulos para realizar la codificación predictiva lineal (LPC), cepstrum derivado de LPC, la predicción lineal perceptiva (PLP), la extracción de características del modelo auditivo y la extracción de características de coeficientes de cepstrum de la frecuencia de Mel (MFCC). Sin embargo, en muchas realizaciones, se utilizan técnicas de extracción de características que producen características cepstrales.

El módulo de extracción de características produce un flujo de vectores de características que están asociados con una trama separada de la señal del habla. Este flujo de vectores de características se proporciona al estimador de la señal limpia 1122.

Las tramas de valores del constructor de tramas 1118 también se proporcionan a un extractor de características 1121, que en una realización extrae la energía de cada trama. El valor de energía para cada trama se proporciona a

una unidad de detección del habla 1126.

En la etapa 1204, la unidad de detección del habla 1126 usa la característica de energía de la señal del sensor alternativo para determinar cuándo es probable que haya habla. Esta información se pasa al entrenador del modelo de ruido 1124, que intenta modelar el ruido durante los períodos en que no hay habla en la etapa 1206.

En una realización, la unidad de detección del habla 1126 primero busca la secuencia de valores de energía de trama para encontrar un pico en la energía. Luego busca un valle después del pico. La energía de este valle se denomina un separador de energía, *d*. Para determinar si una trama contiene habla, la relación, *k*, de la energía de la trama, *e*, sobre el separador de energía, *d*, entonces se determina como: $k=e/d$. Una confianza del habla, *q*, para la trama se determina entonces como:

$$q = \begin{cases} 0 & : \quad k < 1 \\ \dfrac{k-1}{\alpha-1} & : \quad 1 \le k \le \alpha \\ 1 & : \quad k > \alpha \end{cases} \qquad \textbf{EC. 7}$$

donde α define la transición entre dos estados y en una implementación se establece en 2. Finalmente, el valor de confianza promedio de sus 5 tramas vecinas (incluida la propia) se usa como el valor de confianza final para esta trama.

En una realización, se usa un valor umbral fijo para determinar si el habla está presente de manera que, si el valor de confianza sobrepasa el umbral, se considera que la trama contiene habla y si el valor de confianza no sobrepasa el umbral, se considera que la trama no contiene habla En una realización, se usa un valor umbral de 0,1.

Para cada trama sin habla detectada por la unidad de detección del habla 1126, el entrenador del modelo de ruido 1124 actualiza un modelo de ruido 1125 en la etapa 1206. En una realización, el modelo de ruido 1125 es un modelo gaussiano que tiene una media m_n y una varianza $\sum_n$. Este modelo se basa en una ventana móvil de las tramas más recientes sin habla. Las técnicas para determinar la media y la varianza a partir de las tramas sin habla en la ventana son bien conocidas en la técnica.

Los vectores de corrección y los parámetros del modelo en el almacenamiento de parámetros 922 y el modelo de ruido 1125 se proporcionan al estimador de la señal limpia 1122 con los vectores de características, *b*, para el sensor alternativo y los vectores de características, S_y, para la señal ruidosa del micrófono de conducción de aire. En la etapa 1208, el estimador de la señal limpia 1122 estima un valor inicial para la señal del habla limpia basándose en el vector de características del sensor alternativo, los vectores de corrección y los parámetros del modelo para el sensor alternativo. En particular, la estimación del sensor alternativo de la señal limpia se calcula como:

$$\hat{x} = b + \sum_s p(s \mid b) r_s \qquad \textbf{EC. 8}$$

donde $\hat{x}$ es la estimación de la señal limpia en el dominio cepstral, *b* es el vector de características del sensor alternativo, $p(s|b)$ se determina mediante la ecuación 2 anterior, y r_s es el vector de corrección para el componente de la mezcla *s*. Por lo tanto, la estimación de la señal limpia en la ecuación 8 se forma añadiendo el vector de características del sensor alternativo a una suma ponderada de vectores de corrección donde los pesos se basan en la probabilidad de un componente de la mezcla dado el vector de características del sensor alternativo.

En la etapa 1210, la estimación inicial del habla limpia del sensor alternativo se perfecciona combinándola con una estimación del habla limpia que se forma a partir del vector ruidoso del micrófono de conducción de aire y el modelo de ruido. Esto da como resultado una estimación perfeccionada del habla limpia 1128. A fin de combinar el valor cepstral de la estimación inicial de la señal limpia con el vector de características del espectro de potencia del micrófono de conducción de aire ruidoso, el valor cepstral se convierte al dominio del espectro de potencia utilizando:

$$\hat{S}_{x|b} = e^{C^{-1}\hat{x}} \qquad \textbf{EC. 9}$$

donde C^{-1} es una transformada discreta del coseno inversa y $\hat{S}_{x|b}$ es la estimación del espectro de potencia de la señal limpia basada en el sensor alternativo.

Una vez que la estimación inicial de la señal limpia del sensor alternativo se ha colocado en el dominio del espectro de potencia, se puede combinar con el vector ruidoso del micrófono de conducción de aire y el modelo de ruido como:

$$\hat{S}_x = (\Sigma_n^{-1} + \Sigma_{x|b}^{-1})^{-1}[\Sigma_n^{-1}(S_y - \mu_n) + \Sigma_{x|b}^{-1}\hat{S}_{x|b}] \qquad \textbf{EC. 10}$$

donde $\hat{S}_x$ es la estimación perfeccionada de la señal limpia en el dominio del espectro de potencia, S_y es el vector de características ruidoso del micrófono de conducción de aire, $(\mu_n, \sum_n)$ son la media y la covarianza del modelo de ruido anterior (véase 1124), $\hat{S}_{xb}$ es la estimación inicial de la señal limpia basada en el sensor alternativo, y $\sum_{x|b}$ es la matriz de covarianza de la distribución de probabilidad condicional para el habla limpia dada la medición del sensor alternativo. $\sum_{x|b}$ se puede calcular de la siguiente manera. Sea J el jacobiano de la función en el lado derecho de la ecuación 9. Sea $\sum$ la matriz de covarianza de $\hat{x}$. Entonces la covarianza de $\hat{S}_{x|b}$ es

$$\Sigma_{x|b} = J\Sigma J^T \qquad \textbf{EC. 11}$$

En una realización simplificada, EC.10 se reescribe como la ecuación siguiente:

$$\hat{S}_x = \alpha(f)(S_y - \mu_n) + (1-\alpha(f))\hat{S}_{x|b} \qquad \textbf{EC. 12}$$

donde $\alpha(f)$ es una función tanto del tiempo como de la banda de frecuencia. Por ejemplo, si el sensor alternativo tiene un ancho de banda de hasta 3KHz, $\alpha(f)$ se elige para que sea 0 en la banda de frecuencia por debajo de 3KHz. Básicamente, se confía en la estimación inicial de la señal limpia del sensor alternativo para las bandas de frecuencia baja.

En las bandas de alta frecuencia, la estimación inicial de la señal limpia del sensor alternativo no es tan fiable. Intuitivamente, cuando el ruido es pequeño para una banda de frecuencia en la trama actual, se elige un gran $\alpha(f)$ de modo que se coge más información del micrófono de conducción de aire para esta banda de frecuencia. De lo contrario, se utiliza más información del sensor alternativo al elegir un $\alpha(f)$ pequeño. En una realización, la energía de la estimación inicial de la señal limpia del sensor alternativo se usa para determinar el nivel de ruido para cada banda de frecuencia. Sea $E(f)$ la energía para la banda de frecuencia f. Sea $M=Max_f E(f)$. $\alpha(f)$, como una función de f, que se define de la siguiente manera:

$$\alpha(f) = \begin{cases} \dfrac{E(f)}{M} & : \quad f \geq 4K \\ \dfrac{f-3K}{1K}\alpha(4K) & : \quad 3K < f < 4K \\ 0 & : \quad f \leq 3K \end{cases} \qquad \textbf{EC. 13}$$

donde se usa una interpolación lineal para la transición de 3K a 4K para asegurar la suavidad de $\alpha(f)$.

En una realización, la proximidad del dispositivo móvil a la cabeza del usuario se incorpora en la determinación de $\alpha(f)$. Específicamente, si el sensor de proximidad 832 produce un valor de distancia máxima D y un valor de distancia actual d, la ecuación 13 se puede modificar como:

$$\alpha(f) = \begin{cases} \beta\dfrac{E(f)}{M} + (1-\beta)\dfrac{d}{D} & : \quad f \geq 4K \\ \dfrac{f-3K}{1K}\alpha(4K) & : \quad 3K < f < 4K \\ 0 & : \quad f \leq 3K \end{cases} \qquad \textbf{EC. 14}$$

donde β está entre cero y uno y se selecciona en base a qué factor, energía o proximidad, se cree que proporciona la mejor indicación de si el modelo de ruido para el micrófono de conducción de aire o el vector de corrección para el sensor alternativo proporcionará la mejor estimación de la señal limpia.

Si β se establece a cero $\alpha(f)$ ya no depende de la frecuencia y simplemente se convierte en:

$$\alpha = \frac{d}{D} \qquad \textbf{EC. 15}$$

La estimación de la señal limpia perfeccionada en el dominio del espectro de potencia se puede usar para construir un filtro de Wiener y filtrar la señal ruidosa del micrófono de conducción de aire. En particular, el filtro de Wiener, H, se establece de manera que:

$$H = \frac{\hat{S}_x}{S_y} \qquad \textbf{EC. 16}$$

Este filtro se puede aplicar luego contra la señal ruidosa del micrófono de conducción de aire en el dominio del tiempo para producir una señal en el dominio del tiempo limpia o con reducción del ruido. La señal con reducción del ruido se puede proporcionar a un oyente o aplicarse a un reconocedor del habla.

Obsérvese que la ecuación 12 proporciona una estimación de la señal limpia perfeccionada que es la suma ponderada de dos factores, uno de los cuales es una estimación de la señal limpia de un sensor alternativo. Esta suma ponderada se puede ampliar para incluir factores adicionales para sensores alternativos adicionales. Por lo tanto, se puede usar más de un sensor alternativo para generar estimaciones independientes de la señal limpia. Estas estimaciones múltiples se pueden combinar usando la ecuación 12.

En una realización, también se estima el ruido en la estimación de la señal limpia perfeccionada. En una realización, este ruido se trata como una gaussiana de media cero con una covarianza que se determina como:

$$\Sigma_x = (\Sigma_n^{-1} + \Sigma_{x|b}^{-1})^{-1} = \Sigma_n \Sigma_{x|b} /(\Sigma_n + \Sigma_{x|b})$$

donde $\sum_n$ es la variación del ruido en el micrófono de conducción de aire y $\sum_{x|b}$ es la varianza del ruido en la estimación del sensor alternativo. En particular, $\sum_{x|b}$ es más grande si el sensor alternativo no hace buen contacto con la superficie de la piel. La calidad del contacto se puede medir con un sensor de proximidad adicional o analizando el sensor alternativo. Para este último, observando que el sensor alternativo produce poca respuesta de alta frecuencia (mayor que 4KHz) si está en buen contacto, medimos la calidad del contacto con la relación entre la energía de baja frecuencia (menos de 3KHz) y la energía de alta frecuencia. Cuanto mayor sea la relación, mejor será el contacto.

En algunas realizaciones, el ruido en la estimación de la señal limpia se usa para generar un tono lateral como se ha comentado anteriormente en relación con la FIG. 6. A medida que aumenta el ruido en la estimación de la señal limpia perfeccionada, el volumen del tono lateral aumenta para motivar al usuario a colocar el sensor alternativo en una mejor posición de modo que aumente el proceso de mejora. Por ejemplo, el tono lateral motiva a los usuarios a presionar el sensor de conducción ósea contra su cabeza de modo que aumenta el proceso de mejora.

Reducción de ruido mediante el vector de corrección sin la estimación del ruido

La FIG. 13 proporciona un diagrama de bloques de un sistema alternativo para estimar un valor del habla limpia según la presente invención. El sistema de la FIG. 13 es similar al sistema de la FIG. 11 excepto que la estimación del valor del habla limpia se forma sin la necesidad de un micrófono de conducción de aire o un modelo de ruido.

En la FIG. 13, un evento físico asociado con un altavoz 1300 que produce habla se convierte en un vector de características mediante el sensor alternativo 1302, el convertidor analógico-digital 1304, el constructor de tramas 1306 y el extractor de características 1308, de manera similar a la que se comenta anteriormente para el sensor alternativo 1114, convertidor analógico-digital 1116, constructor de tramas 1117 y extractor de características 1118 de la FIG. 11. Obsérvese que, aunque solamente se muestra un sensor alternativo en la FIG. 13, se pueden usar sensores alternativos adicionales como en la FIG. 11 con la incorporación de una unidad de comparación y selección como se ha comentado anteriormente para la FIG. 11.

Los vectores de características del extractor de características 1308 y los parámetros de reducción del ruido 922 se proporcionan a un estimador de la señal limpia 1310, que determina una estimación de un valor de la señal limpia 1312, $\hat{S}_x|b$, usando las ecuaciones 8 y 9 anteriores.

La estimación de la señal limpia, $\hat{S}_{x|b}$, en el dominio del espectro de potencia puede usarse para construir un filtro de Wiener para filtrar una señal ruidosa del micrófono de conducción de aire. En particular, el filtro de Wiener, H, se establece de manera que:

$$H = \frac{\hat{S}_{x|b}}{S_y} \qquad \textbf{EC. 17}$$

Este filtro se puede aplicar luego contra la señal ruidosa del micrófono de conducción de aire en el dominio del tiempo para producir una señal limpia o con reducción del ruido. La señal con reducción del ruido se puede proporcionar a un oyente o aplicarse a un reconocedor del habla.

De forma alternativa, la estimación de la señal limpia en el dominio cepstral, $\hat{x}$, que se calcula en la ecuación 8, puede aplicarse directamente a un sistema de reconocimiento del habla.

Reducción del ruido mediante seguimiento del tono

Una técnica alternativa para generar estimaciones de una señal del habla limpia se muestra en el diagrama de bloques de la FIG. 14 y el diagrama de flujo de la FIG. 15. En particular, la realización de las FIGS. 14 y 15 determina una estimación del habla limpia identificando un tono para la señal del habla usando un sensor alternativo

y luego usando el tono para descomponer una señal ruidosa del micrófono de conducción de aire en un componente armónico y un componente aleatorio. Por lo tanto, la señal ruidosa se representa como:

$$y = y_h + y_r \qquad \text{EC. 18}$$

donde y es la señal ruidosa y_h es el componente armónico, e y_r es el componente aleatorio. Una suma ponderada del componente armónico y el componente aleatorio se utilizan para formar un vector de características con reducción del ruido que representa una señal del habla con reducción del ruido.

En una realización, el componente armónico se modela como una suma de sinusoides relacionados armónicamente de manera que:

$$y_h = \sum_{k=1}^{K} a_k \cos(k\omega_0 t) + b_k \operatorname{sen}(k\omega_0 t) \qquad \text{EC. 19}$$

donde ω_0 es la frecuencia de tono o fundamental y K es el número total de armónicos en la señal.

Por lo tanto, para identificar el componente armónico, se debe determinar una estimación de la frecuencia de tono y los parámetros de amplitud $\{a_1 a_2 ... a_k b_1 b_2 ... b_k\}$.

En la etapa 1500, se recoge una señal del habla ruidosa y se convierte en muestras digitales. Para hacer esto, un micrófono de conducción de aire 1404 convierte las ondas de audio de un altavoz 1400 y una o más fuentes de ruido añadido 1402 en señales eléctricas. Las señales eléctricas son luego muestreadas por un convertidor analógico-digital 1406 para generar una secuencia de valores digitales. En una realización, el convertidor A-D 1406 muestrea la señal analógica a 16 kHz y 16 bits por muestra, creando así 32 kilobytes de datos del habla por segundo. En la etapa 1502, las muestras digitales se agrupan en tramas mediante un constructor de tramas 1408. En una realización, el constructor de tramas 1408 crea una nueva trama cada 10 milisegundos que incluye 25 milisegundos de datos de valor.

En la etapa 1504, un sensor físico 1444 detecta un evento físico asociado con la producción del habla. En esta realización, un sensor alternativo que puede detectar componentes armónicos, como un sensor de conducción ósea, es el más adecuado para ser utilizado como sensor alternativo 1444. Obsérvese que, aunque la etapa 1504 se muestra como separada de la etapa 1500, los expertos en la técnica reconocerán que estas etapas pueden realizarse al mismo tiempo. Además, aunque solamente se muestra un sensor alternativo en la FIG. 14, se pueden usar sensores alternativos adicionales como en la FIG. 11 con la incorporación de una unidad de comparación y selección como se ha comentado anteriormente para la FIG. 11.

La señal analógica generada por el sensor alternativo 1444 se convierte en muestras digitales mediante un convertidor analógico-digital 1446. Las muestras digitales se agrupan en tramas mediante un constructor de tramas 1448 en la etapa 1506.

En la etapa 1508, las tramas de la señal del sensor alternativo son utilizadas por un seguidor de tonos 1450 para identificar el tono o la frecuencia fundamental del habla.

Se puede determinar una estimación de la frecuencia de tono utilizando un número cualquiera de sistemas de seguimiento de tono disponibles. En muchos de estos sistemas, los tonos candidatos se utilizan para identificar posibles espacios entre los centros de segmentos de la señal del sensor alternativo. Para cada tono candidato, se determina una correlación entre segmentos sucesivos del habla. En general, el tono candidato que proporciona la mejor correlación será la frecuencia de tono de la trama. En algunos sistemas, se utiliza información adicional para perfeccionar la selección del tono, como la energía de la señal y/o un seguimiento del tono esperado.

Dada una estimación del tono del seguidor de tonos 1450, el vector de la señal de conducción de aire puede descomponerse en un componente armónico y un componente aleatorio en la etapa 1510. Para hacerlo, la ecuación 19 se reescribe como:

$$\mathbf{y} = \mathbf{A}\mathbf{b} \qquad \text{EC. 20}$$

donde **y** es un vector de N muestras de la señal del habla ruidosa, **A** es una matriz $Nx2K$ dada por:

$$\mathbf{A} = [\mathbf{A}_{\cos} \mathbf{A}_{\operatorname{sen}}] \qquad \text{EC. 21}$$

con elementos

$$\mathbf{A}_{\cos}(k,t) = \cos(k\omega_0 t) \qquad \mathbf{A}_{\operatorname{sen}}(k,t) = \operatorname{sen}(k\omega_0 t) \qquad \text{EC. 22}$$

y b es un vector $2Kx1$ dado por:

$$\mathbf{b}^T = [a_1 a_2 \ldots a_k b_1 b_2 \ldots b_k] \qquad \textbf{EC. 23}$$

Entonces, la solución de mínimos cuadrados para los coeficientes de amplitud es:

$$\hat{\mathbf{b}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \qquad \textbf{EC. 24}$$

Utilizando $\hat{\mathbf{b}}$, una estimación del componente armónico de la señal del habla ruidosa se puede determinar como:

$$\mathbf{y}_h = \mathbf{A}\hat{\mathbf{b}} \qquad \textbf{EC. 25}$$

Una estimación del componente aleatorio se calcula como:

$$\mathbf{y}_r = \mathbf{y} - \mathbf{y}_h \qquad \textbf{EC. 26}$$

Por lo tanto, utilizando las ecuaciones 20-26 anteriores, la unidad de descomposición armónica 1410 puede producir un vector de muestras de componentes armónicos 1412, $\mathbf{y}_h$, y un vector de muestras de componentes aleatorios 1414, $\mathbf{y}_r$.

Después de que las muestras de la trama se hayan descompuesto en muestras armónicas y aleatorias, se determina un parámetro de escalado o peso para el componente armónico en la etapa 1512. Este parámetro de escalado se utiliza como parte de un cálculo de una señal del habla con reducción del ruido como se describe más adelante. En una realización, el parámetro de escalado se calcula como:

$$\alpha_h = \frac{\sum_i y_h(i)^2}{\sum_i y(i)^2} \qquad \textbf{EC. 27}$$

donde α_h es el parámetro de escalado, $y_h(i)$ es la i-ésima muestra en el vector de muestras de componentes armónicos $\mathbf{y}_h$ e $y(i)$ es la i-ésima muestra de la señal del habla ruidosa para esta trama. En la ecuación 27, el numerador es la suma de la energía de cada muestra del componente armónico y el denominador es la suma de la energía de cada muestra de la señal del habla ruidosa. Por lo tanto, el parámetro de escalado es la relación entre la energía armónica de la trama y la energía total de la trama.

En realizaciones alternativas, el parámetro de escalado se establece usando una unidad de detección probabilística sonora-sorda. Dichas unidades brindan la probabilidad de que se exprese una trama particular del habla, lo que significa que las cuerdas vocales resuenan durante la trama, en lugar de ser sordas. La probabilidad de que la trama sea de una región del habla sonora se puede usar directamente como parámetro de escalado.

Después de que se haya determinado el parámetro de escalado o mientras se está determinando, los espectros de Mel para el vector de muestras de componentes armónicos y el vector de muestras de componentes aleatorios se determinan en la etapa 1514. Esto implica pasar cada vector de muestras a través de una transformada discreta de Fourier (DFT) 1418 para producir un vector de valores de frecuencia de componentes armónicos 1422 y un vector de valores de frecuencia de componentes aleatorios 1420. Los espectros de potencia representados por los vectores de valores de frecuencia son luego suavizados por una unidad de ponderación de Mel 1424 usando una serie de funciones de ponderación triangular aplicadas a lo largo de la escala de Mel. Esto da como resultado un vector espectral de componentes armónicos de Mel 1428, $\mathbf{Y}_h$, y un vector espectral de componentes aleatorios de Mel 1426, $\mathbf{Y}_r$.

En la etapa 1516, los espectros de Mel para el componente armónico y el componente aleatorio se combinan como una suma ponderada para formar una estimación de un espectro de Mel con reducción del ruido. Esta etapa se realiza mediante la calculadora de la suma ponderada 1430 utilizando el factor de escalado determinado anteriormente en la siguiente ecuación:

$$\hat{\mathbf{X}}(t) = \alpha_h(t)\mathbf{Y}_h(t) + \alpha_r\mathbf{Y}_r(t) \qquad \textbf{EC. 28}$$

donde $\hat{\mathbf{X}}(t)$ es la estimación del espectro de Mel con reducción del ruido, $\mathbf{Y}_h(t)$ es el componente armónico del espectro de Mel, $\mathbf{Y}_r(t)$ es el componente aleatorio del espectro de Mel, $\alpha_h(t)$ es el factor de escalado determinado anteriormente, α_r es un factor de escalado fijo para el componente aleatorio que en una realización se establece igual a 0,1, y el índice de tiempo t se usa para enfatizar que el factor de escalado para el componente armónico se determina para cada trama mientras que el factor de escalado para el componente aleatorio permanece fijo. Obsérvese que, en otras realizaciones, el factor de escalado para el componente aleatorio se puede determinar para cada trama.

Después de calcular el espectro de Mel con reducción del ruido en la etapa 1516, se determina el logaritmo 1432 del espectro de Mel y luego se aplica a una transformada discreta del coseno 1434 en la etapa 1518. Esto produce un

vector de características de coeficientes cepstrales de frecuencia de Mel (MFCC) 1436 que representa una señal del habla con reducción del ruido.

Se produce un vector de características MFCC con reducción del ruido separado para cada trama de la señal ruidosa. Estos vectores de características pueden usarse para cualquier propósito deseado, incluyendo mejora del habla y reconocimiento del habla. Para mejorar el habla, los vectores de características de MFCC se pueden convertir en el dominio del espectro de potencia y se pueden usar con la señal de conducción de aire ruidosa para formar un filtro de Weiner.

Aunque la presente invención se ha analizado anteriormente con referencia específica al uso de sensores de conducción ósea como sensores alternativos, se pueden usar otros sensores alternativos. Por ejemplo, en la FIG. 16, un dispositivo móvil de la presente invención utiliza un sensor infrarrojo 1600 que en general está dirigido a la cara del usuario, especialmente la región de la boca, y genera una señal indicativa de un cambio en el movimiento facial del usuario que corresponde al habla. La señal generada por el sensor infrarrojo 1600 se puede usar como la señal del sensor alternativo en las técnicas descritas anteriormente.

REIVINDICACIONES

1. Un dispositivo móvil de mano, que comprende:

un micrófono de conducción de aire (108) que está configurado para convertir ondas acústicas en una señal de micrófono eléctrica,

al menos uno de entre un primer sensor alternativo y un segundo sensor alternativo, dicho primer sensor alternativo distinto del micrófono de conducción de aire que está configurado para proporcionar una primera señal eléctrica del sensor alternativo indicativa del habla,

dicho segundo sensor alternativo distinto del micrófono de conducción de aire que está configurado para proporcionar una segunda señal del sensor alternativo, y

un procesador que está configurado para usar la señal del micrófono y la primera y segunda señal de sensor alternativo para estimar un valor del habla limpia,

en el que el primer sensor alternativo y el segundo sensor alternativo comprenden sensores de conducción ósea, el dispositivo móvil de mano incluye un altavoz que está configurado para colocarse cerca del oído izquierdo o derecho del usuario, el altavoz y el primer y el segundo sensor alternativo miran hacia la misma dirección, y cuando el dispositivo está colocado cerca del oído izquierdo o derecho del usuario y el altavoz mira hacia el oído respectivo, el primer o el segundo sensor de conducción ósea está configurado para entrar en contacto con el cráneo o el oído del usuario y producir la señal del sensor alternativo.

2. El dispositivo móvil de mano de la reivindicación 1, en el que el dispositivo móvil de mano incluye una pantalla ubicada debajo del altavoz, y cuando el altavoz mira hacia el usuario, el dispositivo móvil de mano tiene un lado izquierdo y un lado derecho opuesto al lado izquierdo y en el que el primer sensor alternativo está ubicado cerca del lado izquierdo y el segundo sensor alternativo está ubicado cerca del lado derecho.

3. El dispositivo móvil de mano de la reivindicación 1, que comprende además una unidad de selección que selecciona una de la primera señal de sensor alternativo y la segunda señal de sensor alternativo.

4. El dispositivo móvil de mano de la reivindicación 3, en el que la unidad de selección está configurada para seleccionar una de la primera señal de sensor alternativo y la segunda señal de sensor alternativo en base a las magnitudes de la primera señal de sensor alternativo y la segunda señal de sensor alternativo.

5. El dispositivo móvil de mano de la reivindicación 1, en el que el altavoz (110) está configurado para generar un sonido basado en la cantidad de ruido en el valor del habla limpia.

6. El dispositivo móvil de mano de la reivindicación 1, que comprende además un sensor de proximidad que está configurado para producir una señal de proximidad indicativa de la distancia entre el dispositivo móvil de mano y un objeto.

7. El dispositivo móvil de mano de la reivindicación 6, en el que el procesador está configurado para determinar el valor del habla limpia basándose en la señal del micrófono, la primera y/o segunda señal de sensor alternativo y la señal de proximidad.

8. El dispositivo móvil de mano de la reivindicación 7, en el que el procesador está configurado para determinar el valor del habla limpia a través de un proceso que comprende:

determinar una contribución del micrófono al valor del habla limpia en base a la señal del micrófono;

determinar una contribución de sensor alternativo al valor del habla limpia en base a la primera y/o segunda señal de sensor alternativo; y

ponderar la contribución del micrófono y la contribución del sensor alternativo en base a la señal de proximidad.

9. El dispositivo móvil de mano de la reivindicación 6, en el que el altavoz (110) está configurado para generar un sonido basado en la señal de proximidad.

10. Un dispositivo móvil, que comprende:

un micrófono de conducción de aire (508) que está configurado para convertir ondas acústicas en una señal de micrófono eléctrica;

un sensor alternativo (512) distinto del micrófono de conducción de aire que está configurado para proporcionar una señal eléctrica de sensor alternativo indicativa del habla;

un sensor de proximidad que está configurado para proporcionar una señal de proximidad eléctrica que es indicativa de la distancia desde el dispositivo móvil a un objeto; y

un estimador de la señal limpia que está configurado para usar la señal del micrófono, la señal del sensor alternativo y la señal de proximidad para eliminar el ruido de la señal del micrófono y, por lo tanto, producir una señal del habla limpia mejorada,

en el que el dispositivo móvil produce la señal del habla limpia mejorada utilizando la señal de proximidad para ponderar una contribución a la señal del habla limpia mejorada que está formada a partir de la señal del sensor alternativo.

11. El dispositivo móvil de la reivindicación 10, que comprende además un altavoz (510) que está configurado para producir un sonido basado en una estimación del nivel de ruido en la señal del habla limpia mejorada.

12. El dispositivo móvil de la reivindicación 10, en el que ponderar la contribución comprende dar menos peso a la contribución cuando la señal de proximidad indica que el dispositivo móvil está lejos del objeto.

13. El dispositivo móvil de la reivindicación 10, que comprende además un altavoz (510) que está configurado para producir un sonido basado en la señal de proximidad.

14. El dispositivo móvil de la reivindicación 13, en el que el volumen del sonido está configurado para aumentar a medida que la señal de proximidad indica que la distancia entre el dispositivo móvil y el objeto aumenta.

15. El dispositivo móvil de la reivindicación 10, en el que la señal del sensor alternativo y la señal del sensor de proximidad son producidas por un único sensor.

16. El dispositivo móvil de la reivindicación 15, en el que el sensor único comprende un transductor de presión (604) que proporciona una señal eléctrica, la señal eléctrica que tiene un componente de CC que representa la señal de proximidad y un componente de CA que representa la señal del sensor alternativo.

17. Un método en un dispositivo móvil, el método que comprende:

recibir una señal del micrófono de conducción de aire;

recibir una señal del sensor alternativo que es indicativa del habla desde un sensor alternativo distinto del micrófono de conducción de aire;

estimar un valor del habla limpia mejorada basándose en la señal del micrófono de conducción de aire y la señal del sensor alternativo;

estimar el ruido en el valor del habla limpia mejorada; y

usar la estimación del ruido para generar un sonido a través de un altavoz en el dispositivo móvil; y

que comprende además recibir una señal del sensor de proximidad que indica la distancia entre el dispositivo móvil y un objeto y usar la señal del sensor de proximidad cuando se estima el valor del habla limpia mejorada,

en el que el uso de la señal del sensor de proximidad comprende ponderar una contribución al valor del habla limpia mejorada que se deriva de la señal del sensor alternativo basándose en la señal del sensor de proximidad.

18. El método de la reivindicación 17, en el que el volumen del sonido está configurado para aumentar a medida que aumenta la estimación del ruido.

19. El método de la reivindicación 17, en el que ponderar una contribución comprende aplicar un mayor peso a la contribución derivada de la señal del sensor alternativo cuando la señal del sensor de proximidad indica que el dispositivo móvil está cerca de un objeto.

20. El método de la reivindicación 19, en el que el objeto es la cabeza de un usuario.

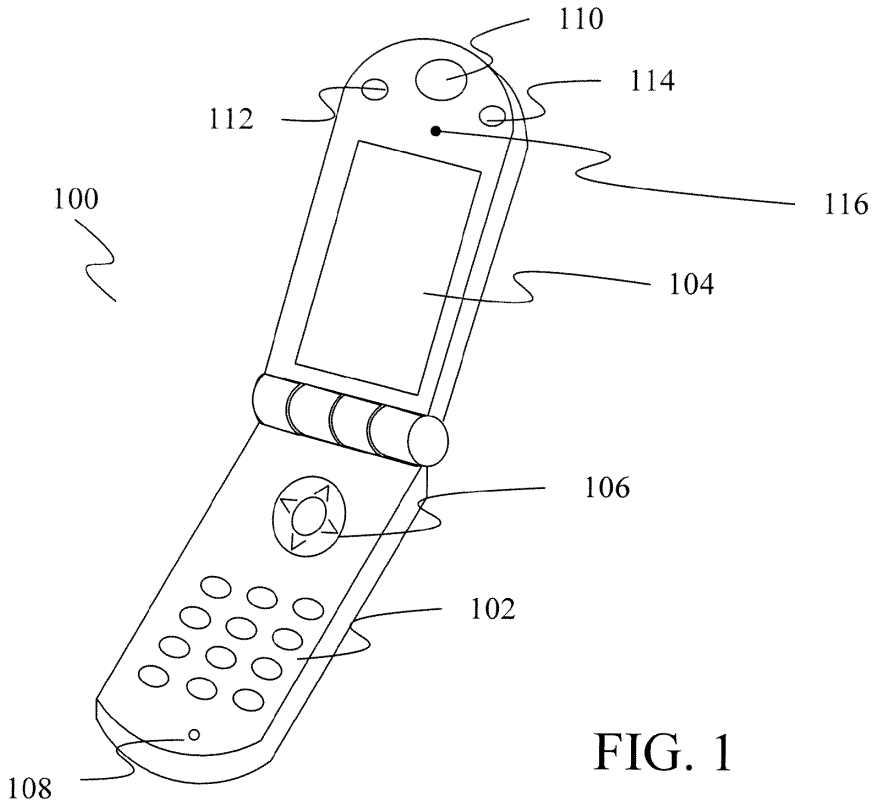

FIG. 1

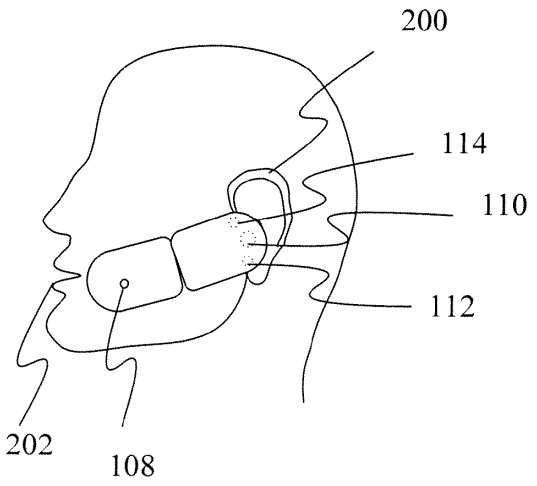

FIG. 2

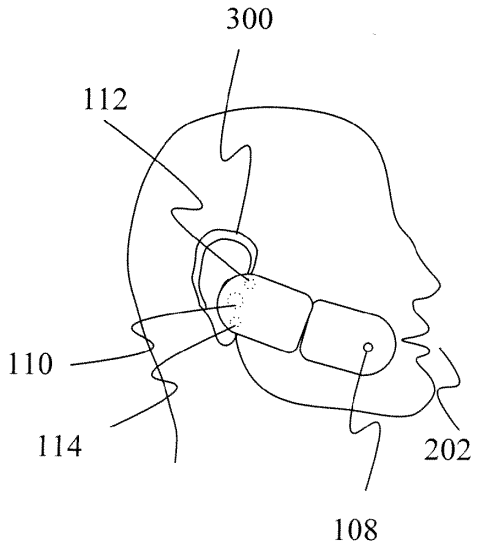

FIG. 3

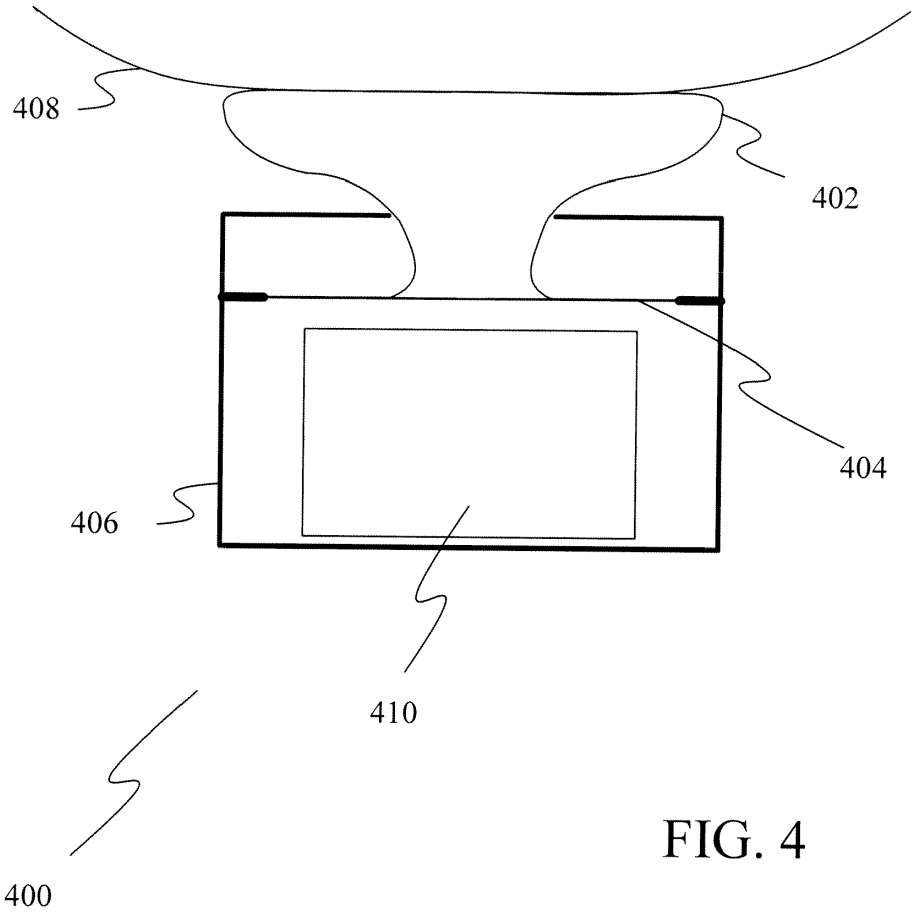

FIG. 4

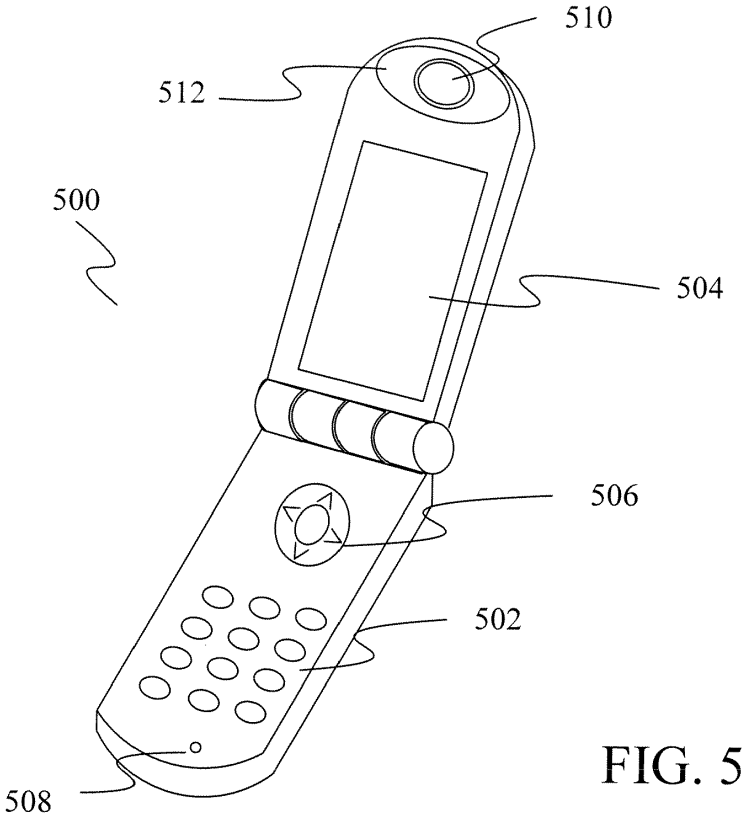

FIG. 5

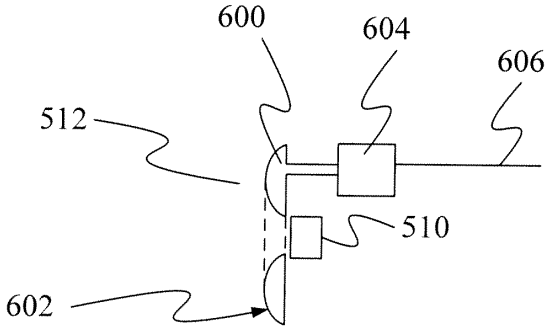

FIG. 6

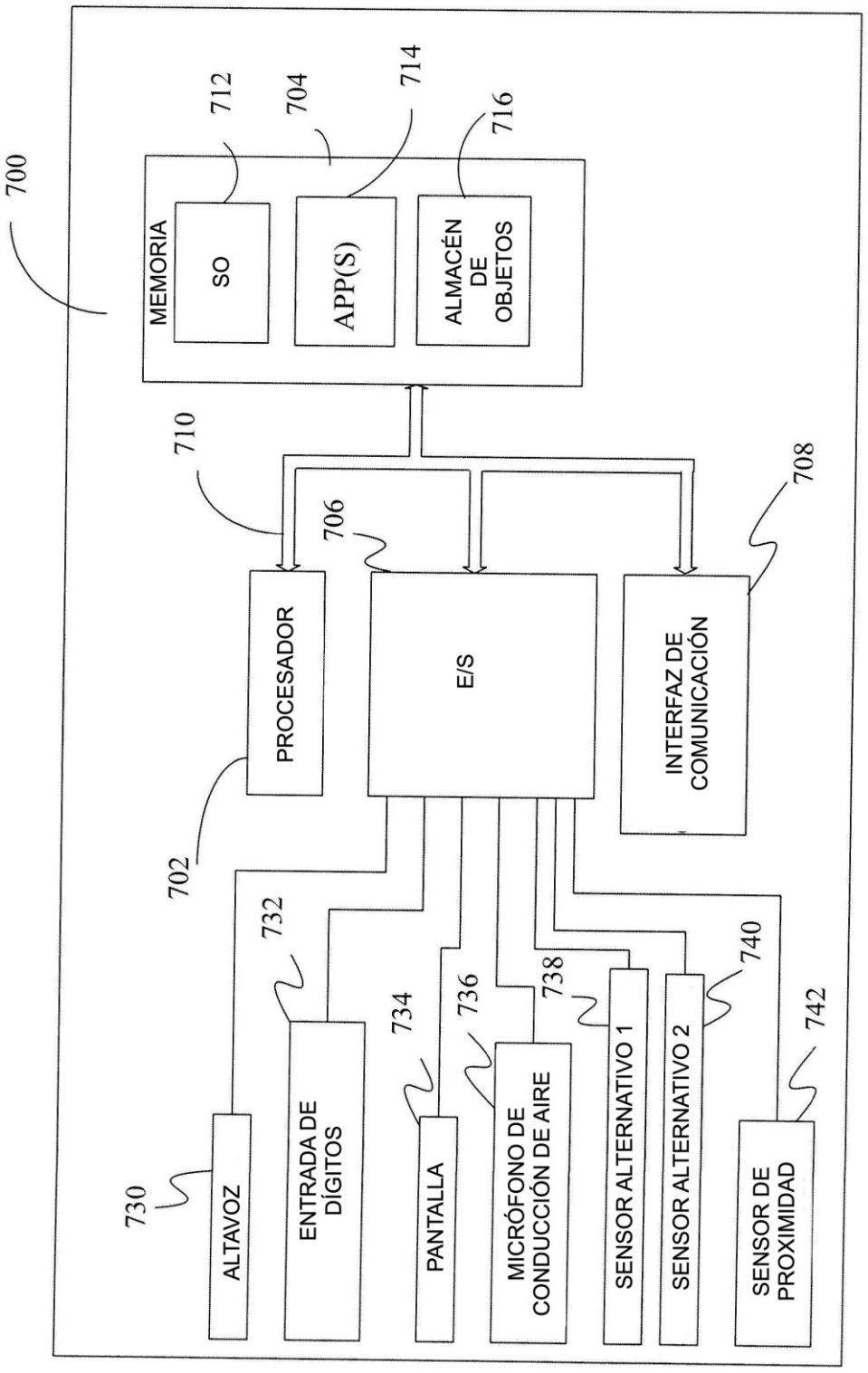
700
MEMORIA
SO
712
704
APP(S)
714
ALMACÉN DE OBJETOS
716
710
706
708
PROCESADOR
E/S
INTERFAZ DE COMUNICACIÓN
702
732
734
736
738
740
742
730
ALTAVOZ
ENTRADA DE DÍGITOS
PANTALLA
MICRÓFONO DE CONDUCCIÓN DE AIRE
SENSOR ALTERNATIVO 1
SENSOR ALTERNATIVO 2
SENSOR DE PROXIMIDAD

FIG. 7

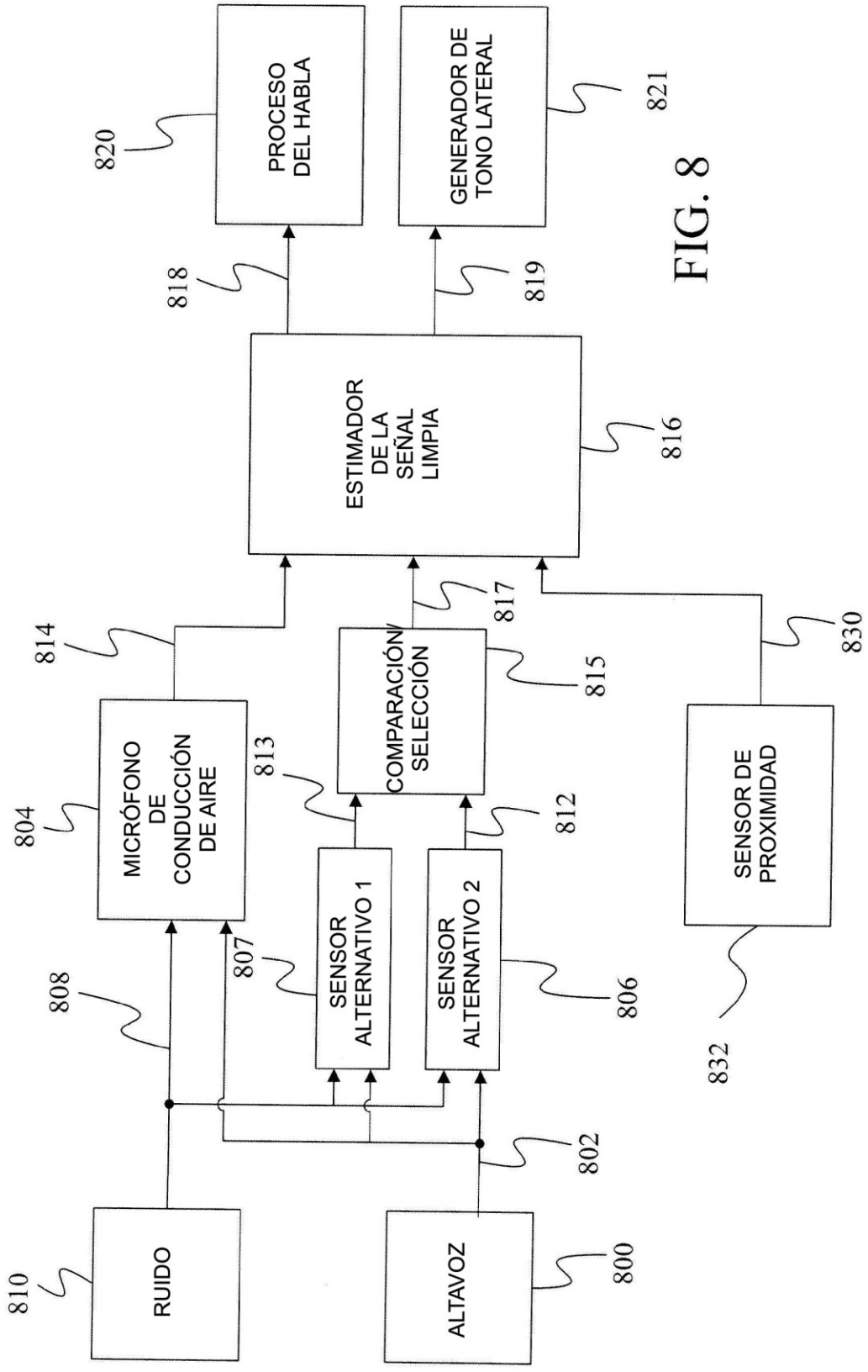
PROCESO DEL HABLA
GENERADOR DE TONO LATERAL
ESTIMADOR DE LA SEÑAL LIMPIA
COMPARACIÓN SELECCIÓN
MICRÓFONO DE CONDUCCIÓN DE AIRE
SENSOR ALTERNATIVO 1
SENSOR ALTERNATIVO 2
SENSOR DE PROXIMIDAD
RUIDO
ALTAVOZ
820
821
818
819
816
817
814
815
813
812
804
807
808
806
832
830
802
810
800

FIG. 8

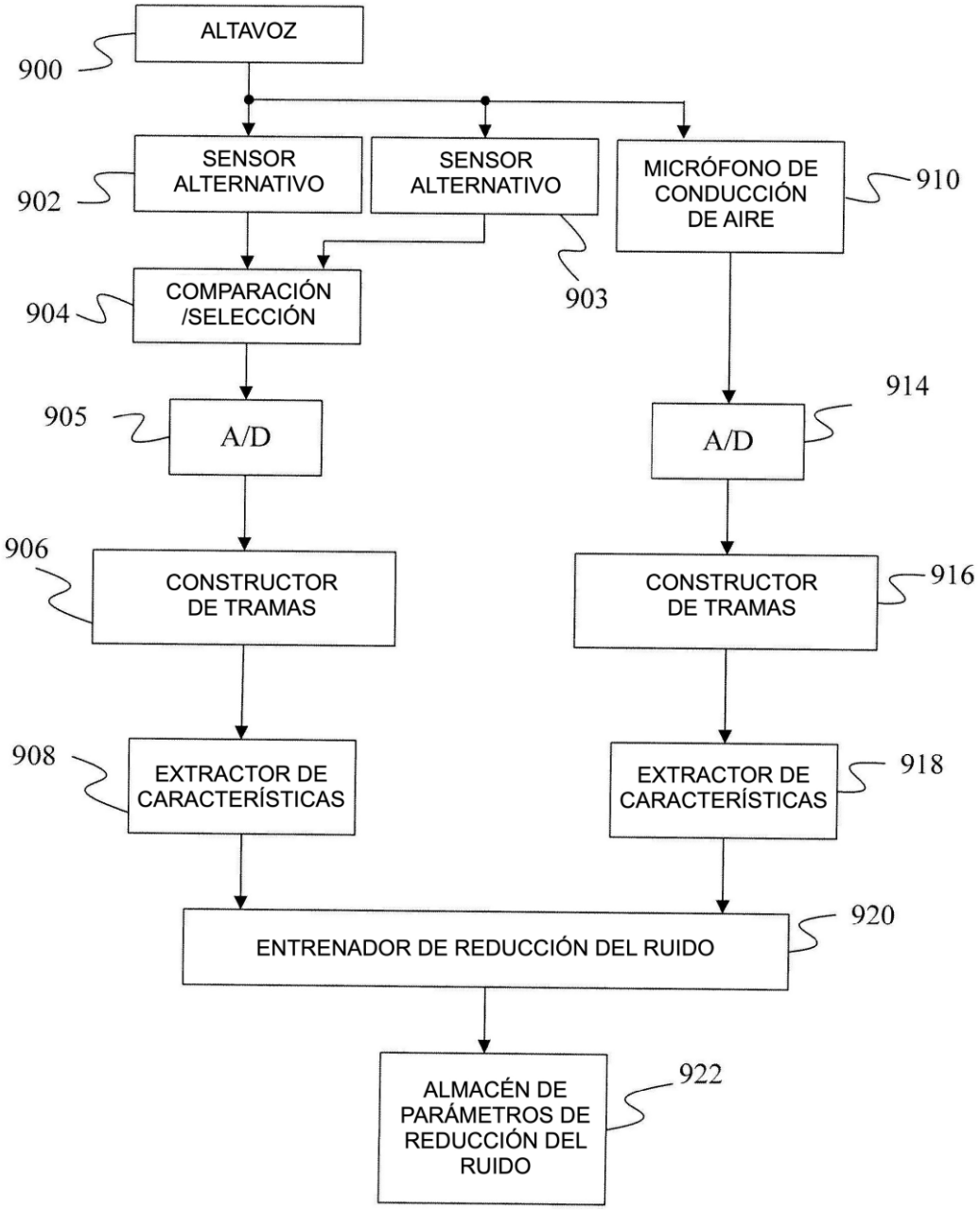
ALTAVOZ
900
SENSOR ALTERNATIVO
902
SENSOR ALTERNATIVO
MICRÓFONO DE CONDUCCIÓN DE AIRE
910
903
COMPARACIÓN /SELECCIÓN
904
905
A/D
914
A/D
906
CONSTRUCTOR DE TRAMAS
CONSTRUCTOR DE TRAMAS
916
908
EXTRACTOR DE CARACTERÍSTICAS
EXTRACTOR DE CARACTERÍSTICAS
918
ENTRENADOR DE REDUCCIÓN DEL RUIDO
920
ALMACÉN DE PARÁMETROS DE REDUCCIÓN DEL RUIDO
922

FIG. 9

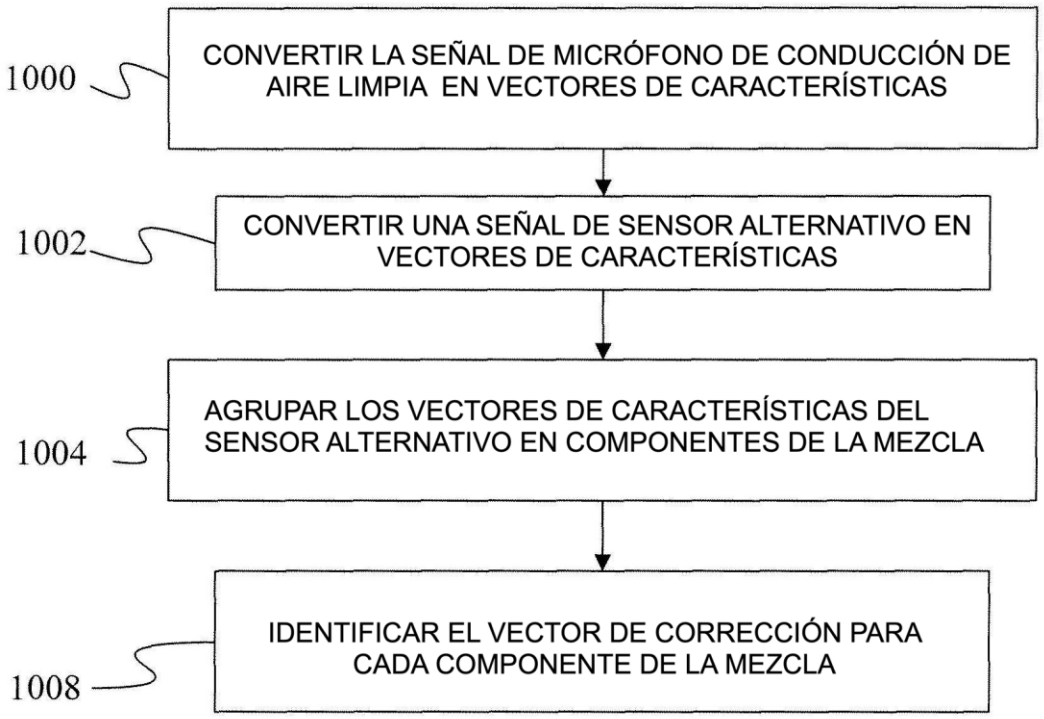
1000
CONVERTIR LA SEÑAL DE MICRÓFONO DE CONDUCCIÓN DE AIRE LIMPIA EN VECTORES DE CARACTERÍSTICAS
1002
CONVERTIR UNA SEÑAL DE SENSOR ALTERNATIVO EN VECTORES DE CARACTERÍSTICAS
1004
AGRUPAR LOS VECTORES DE CARACTERÍSTICAS DEL SENSOR ALTERNATIVO EN COMPONENTES DE LA MEZCLA
1008
IDENTIFICAR EL VECTOR DE CORRECCIÓN PARA CADA COMPONENTE DE LA MEZCLA

FIG. 10

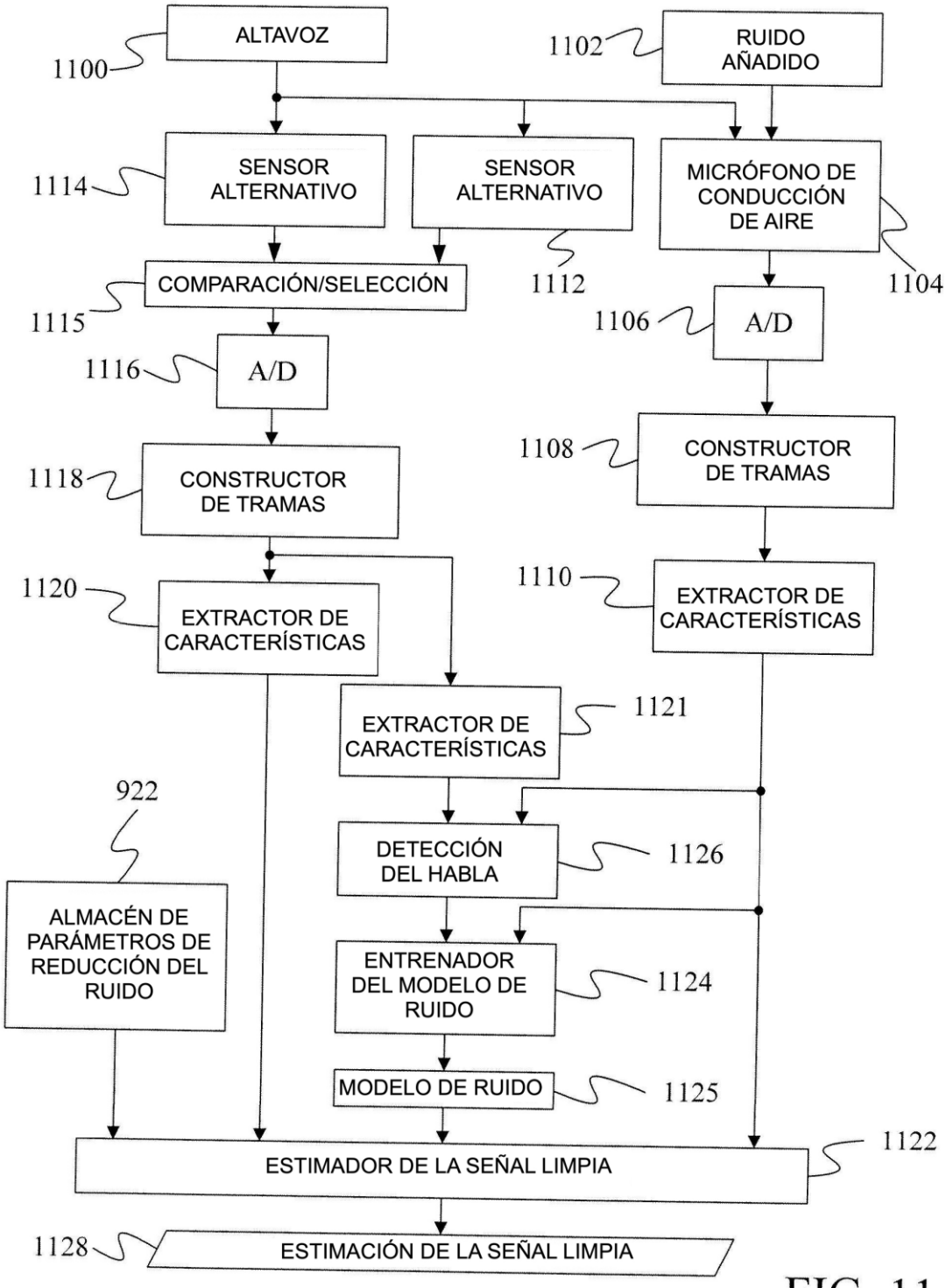
ALTAVOZ
1100
1102
RUIDO AÑADIDO
SENSOR ALTERNATIVO
1114
SENSOR ALTERNATIVO
1112
MICRÓFONO DE CONDUCCIÓN DE AIRE
1104
COMPARACIÓN/SELECCIÓN
1115
A/D
1116
1106
A/D
CONSTRUCTOR DE TRAMAS
1118
1108
CONSTRUCTOR DE TRAMAS
EXTRACTOR DE CARACTERÍSTICAS
1120
1110
EXTRACTOR DE CARACTERÍSTICAS
EXTRACTOR DE CARACTERÍSTICAS
1121
922
DETECCIÓN DEL HABLA
1126
ALMACÉN DE PARÁMETROS DE REDUCCIÓN DEL RUIDO
ENTRENADOR DEL MODELO DE RUIDO
1124
MODELO DE RUIDO
1125
ESTIMADOR DE LA SEÑAL LIMPIA
1122
1128
ESTIMACIÓN DE LA SEÑAL LIMPIA

FIG. 11

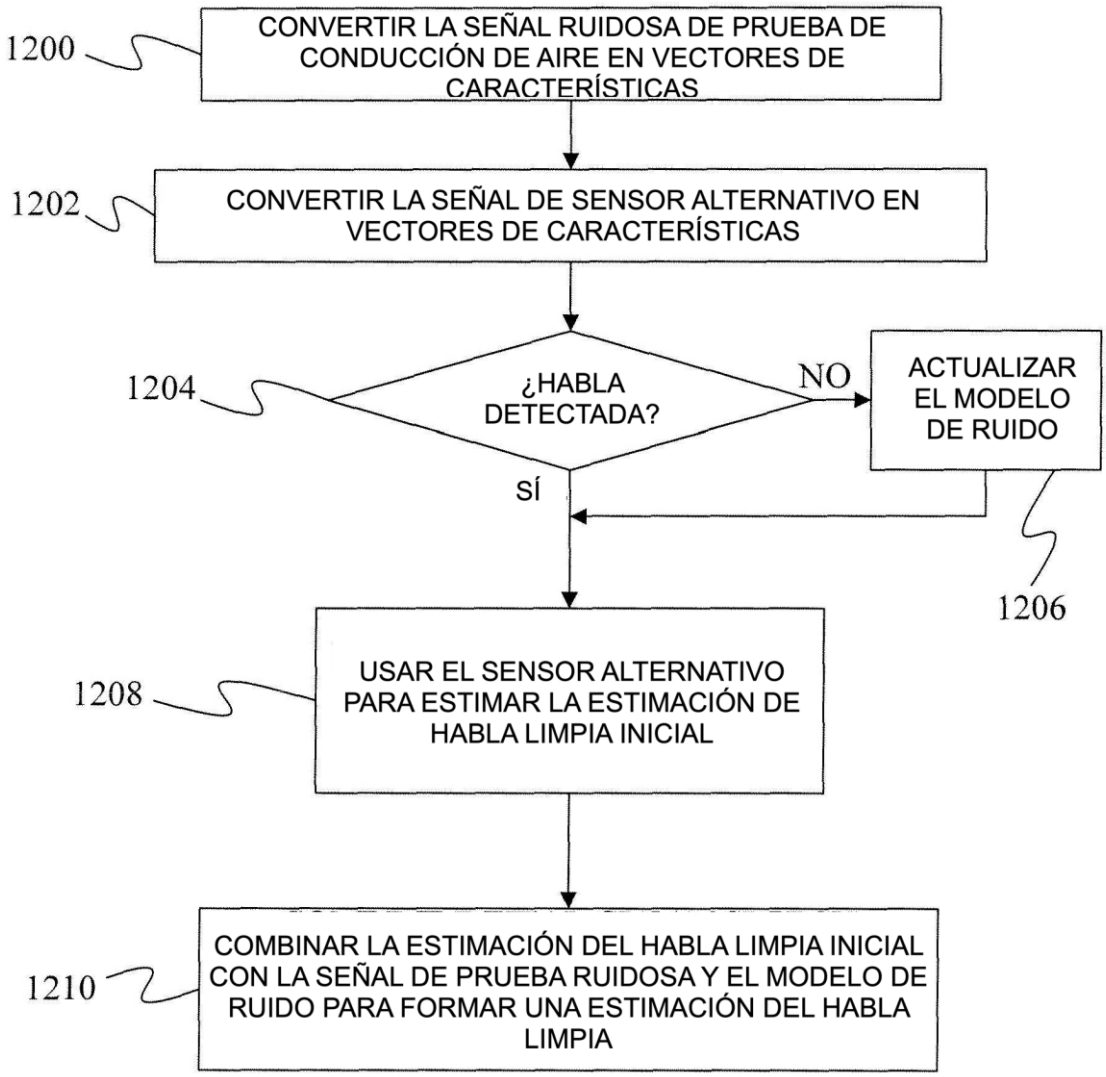
1200
CONVERTIR LA SEÑAL RUIDOSA DE PRUEBA DE CONDUCCIÓN DE AIRE EN VECTORES DE CARACTERÍSTICAS
1202
CONVERTIR LA SEÑAL DE SENSOR ALTERNATIVO EN VECTORES DE CARACTERÍSTICAS
1204
¿HABLA DETECTADA?
NO
ACTUALIZAR EL MODELO DE RUIDO
SÍ
1206
1208
USAR EL SENSOR ALTERNATIVO PARA ESTIMAR LA ESTIMACIÓN DE HABLA LIMPIA INICIAL
1210
COMBINAR LA ESTIMACIÓN DEL HABLA LIMPIA INICIAL CON LA SEÑAL DE PRUEBA RUIDOSA Y EL MODELO DE RUIDO PARA FORMAR UNA ESTIMACIÓN DEL HABLA LIMPIA

FIG. 12

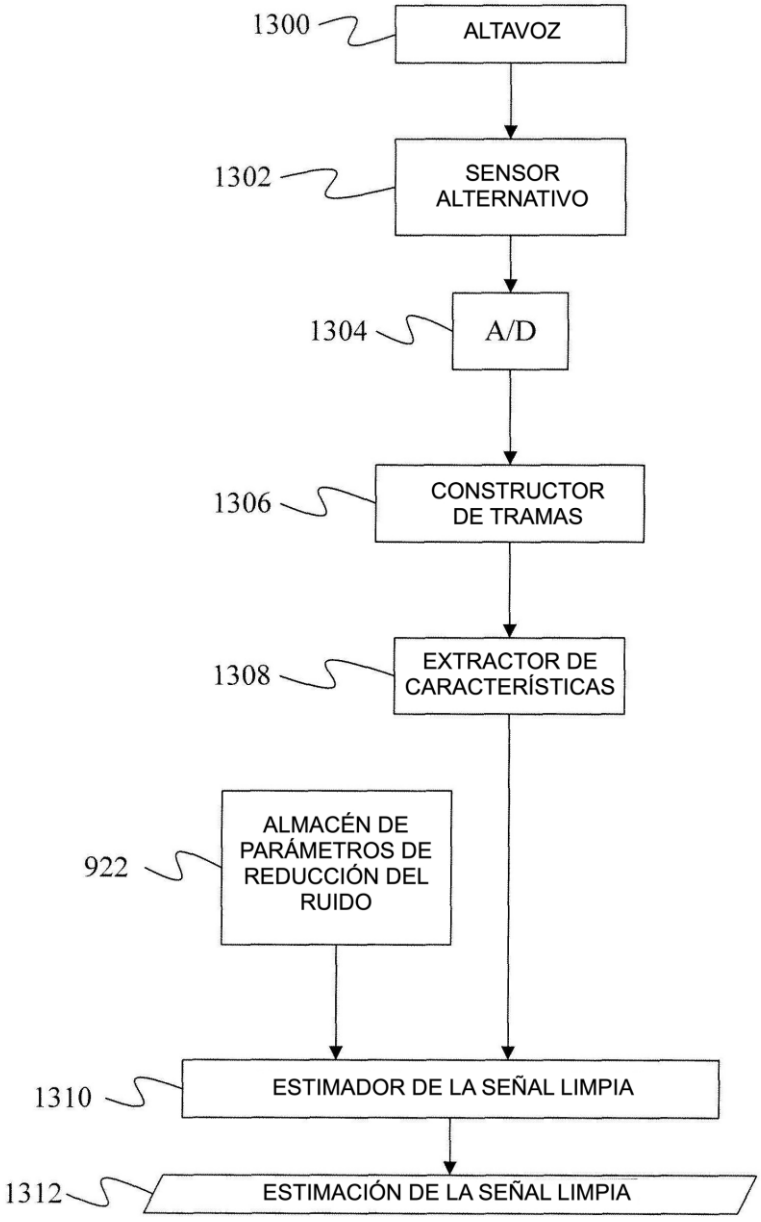
1300
ALTAVOZ
1302
SENSOR ALTERNATIVO
1304
A/D
1306
CONSTRUCTOR DE TRAMAS
1308
EXTRACTOR DE CARACTERÍSTICAS
922
ALMACÉN DE PARÁMETROS DE REDUCCIÓN DEL RUIDO
1310
ESTIMADOR DE LA SEÑAL LIMPIA
1312
ESTIMACIÓN DE LA SEÑAL LIMPIA

FIG. 13

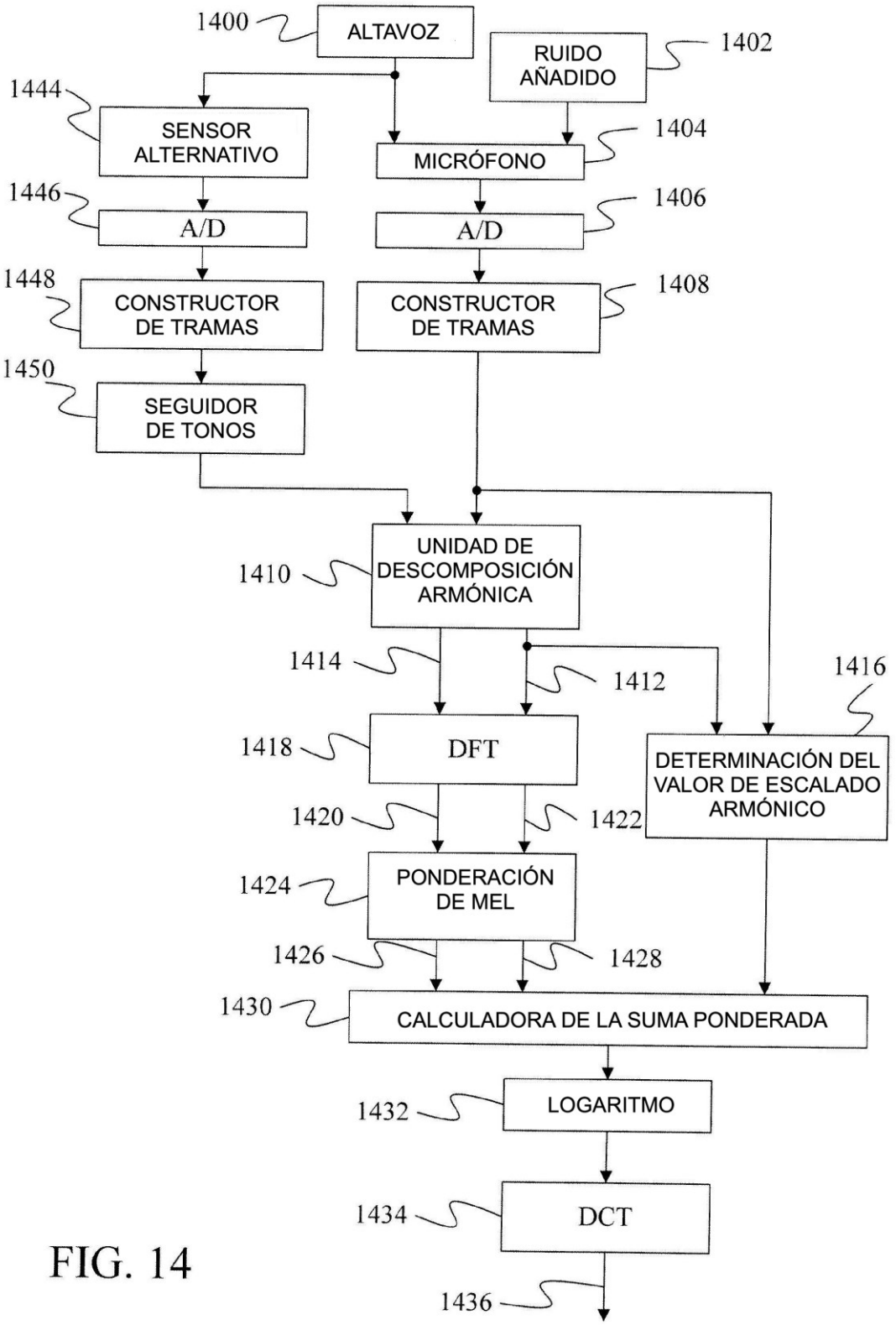
1400
ALTAVOZ
RUIDO AÑADIDO
1402
1444
SENSOR ALTERNATIVO
MICRÓFONO
1404
1446
A/D
A/D
1406
1448
CONSTRUCTOR DE TRAMAS
CONSTRUCTOR DE TRAMAS
1408
1450
SEGUIDOR DE TONOS
UNIDAD DE DESCOMPOSICIÓN ARMÓNICA
1410
1414
1412
1416
1418
DFT
DETERMINACIÓN DEL VALOR DE ESCALADO ARMÓNICO
1420
1422
1424
PONDERACIÓN DE MEL
1426
1428
1430
CALCULADORA DE LA SUMA PONDERADA
1432
LOGARITMO
1434
DCT
1436

FIG. 14

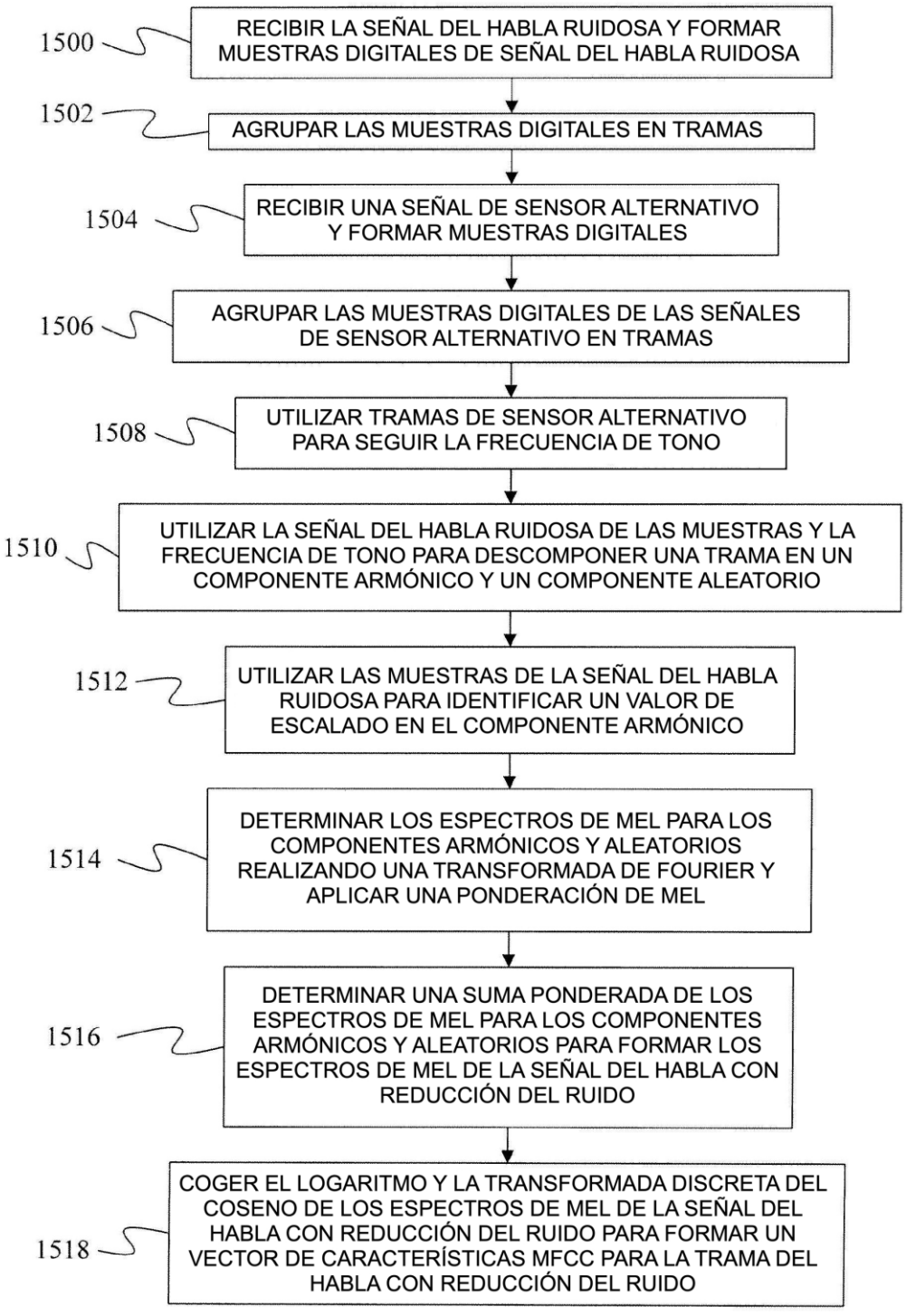
1500
RECIBIR LA SEÑAL DEL HABLA RUIDOSA Y FORMAR MUESTRAS DIGITALES DE SEÑAL DEL HABLA RUIDOSA
1502
AGRUPAR LAS MUESTRAS DIGITALES EN TRAMAS
1504
RECIBIR UNA SEÑAL DE SENSOR ALTERNATIVO Y FORMAR MUESTRAS DIGITALES
1506
AGRUPAR LAS MUESTRAS DIGITALES DE LAS SEÑALES DE SENSOR ALTERNATIVO EN TRAMAS
1508
UTILIZAR TRAMAS DE SENSOR ALTERNATIVO PARA SEGUIR LA FRECUENCIA DE TONO
1510
UTILIZAR LA SEÑAL DEL HABLA RUIDOSA DE LAS MUESTRAS Y LA FRECUENCIA DE TONO PARA DESCOMPONER UNA TRAMA EN UN COMPONENTE ARMÓNICO Y UN COMPONENTE ALEATORIO
1512
UTILIZAR LAS MUESTRAS DE LA SEÑAL DEL HABLA RUIDOSA PARA IDENTIFICAR UN VALOR DE ESCALADO EN EL COMPONENTE ARMÓNICO
1514
DETERMINAR LOS ESPECTROS DE MEL PARA LOS COMPONENTES ARMÓNICOS Y ALEATORIOS REALIZANDO UNA TRANSFORMADA DE FOURIER Y APLICAR UNA PONDERACIÓN DE MEL
1516
DETERMINAR UNA SUMA PONDERADA DE LOS ESPECTROS DE MEL PARA LOS COMPONENTES ARMÓNICOS Y ALEATORIOS PARA FORMAR LOS ESPECTROS DE MEL DE LA SEÑAL DEL HABLA CON REDUCCIÓN DEL RUIDO
1518
COGER EL LOGARITMO Y LA TRANSFORMADA DISCRETA DEL COSENO DE LOS ESPECTROS DE MEL DE LA SEÑAL DEL HABLA CON REDUCCIÓN DEL RUIDO PARA FORMAR UN VECTOR DE CARACTERÍSTICAS MFCC PARA LA TRAMA DEL HABLA CON REDUCCIÓN DEL RUIDO

FIG. 15

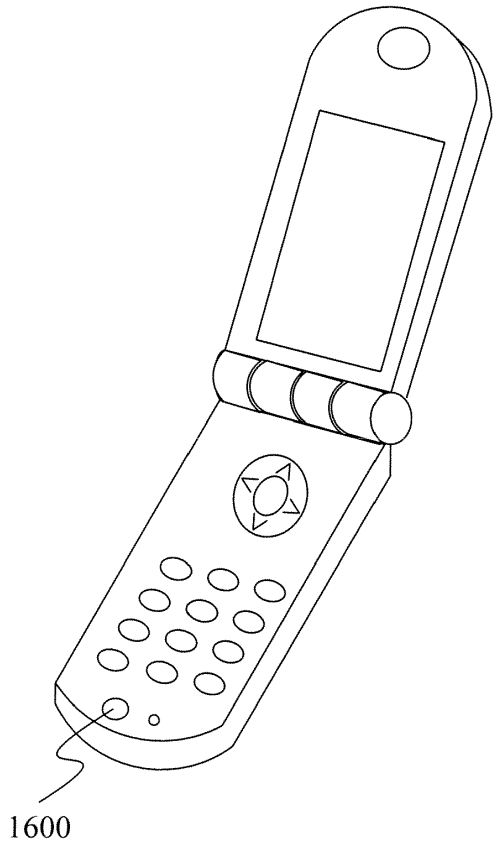

FIG. 16