(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 661 924**

(51) Int. CI.:

G10L 25/78 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **30.08.2013 E 16184741 (3)**

(97) Fecha y número de publicación de la concesión europea: **06.12.2017 EP 3113184**

(54) Título: **Método y dispositivo para detectar la actividad vocal**

(30) Prioridad:

31.08.2012 US 201261695623 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
04.04.2018

(73) Titular/es:

TELEFONAKTIEBOLAGET LM ERICSSON (PUBL) (100.0%)
164 83 Stockholm, SE

(72) Inventor/es:

SEHLSTEDT, MARTIN

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Método y dispositivo para detectar la actividad vocal

Campo técnico

La presente descripción se refiere en general a un método y a un dispositivo para detectar la actividad vocal (VAD).

Antecedentes

En los sistemas de codificación de voz utilizados para el lenguaje en conversación es corriente utilizar la transmisión discontinua (DTX) para aumentar el rendimiento de la codificación. La razón es que el lenguaje en conversación contiene gran cantidad de pausas incrustadas en la conversación, es decir, mientras una persona está hablando, la otra está escuchando. Así, con DTX, el codificador de voz está solo activo alrededor del 50% del tiempo medio y el resto se puede codificar utilizando ruido de confort. Algunos códecs de ejemplo que disponen de esta característica son Banda Estrecha Adaptativa Multi Velocidad (AMR NB) y Códec Mejorado de Velocidad Variable (EVRC). AMR NB utiliza DTX y EVRC utiliza velocidad variable de bits (VBR), en los que un Algoritmo de Determinación de la Velocidad (RDA) decide qué velocidad de datos utilizar para cada trama, basándose en una decisión VAD. En la operación DTX, las tramas activas de voz se codifican utilizando el códec mientras que las tramas entre regiones activas ser reemplazan por el ruido confortable. Los parámetros del ruido confortable se estiman en el codificador y se envían al descodificador utilizando una velocidad de trama reducida y una velocidad de bits menor que la utilizada para la conversación activa.

Para el funcionamiento en DTX de alta calidad, es decir, sin degradación de la calidad de la voz, es importante detectar los períodos de conversación en la señal de entrada. Esto se realiza normalmente por medio del detector de actividad vocal (VAD) (que se usa tanto en DTX como en RDA). La figura 1 muestra un diagrama general de bloques de ejemplo de un VAD 100 generalizado, el cual toma la señal de entrada 111, dividida típicamente en tramas de datos de 5-30 ms dependiendo de la ejecución, como entrada y produce decisiones VAD como salida, normalmente una decisión para cada trama. Es decir, una decisión VAD es una decisión para cada trama si la trama contiene voz o ruido.

La decisión preliminar, vad_prim 113, la toma en este ejemplo el detector primario de voz 101 y es justamente en este ejemplo básicamente una comparación de las características para la trama actual y las características ambientales (estimadas normalmente a partir de las tramas previas de entrada), en la que una diferencia mayor de un cierto umbral genera una decisión primaria activa. En otros ejemplos, la decisión preliminar se puede lograr de otros modos, algunos de los cuales se desarrollan brevemente más adelante. Los detalles del funcionamiento interno del detector primario de voz no es de crucial importancia para la presente descripción y cualquier detector primario de voz que produzca una decisión preliminar será útil en el presente contexto. El bloque de adición del tiempo de espera 102 se usa en el presente ejemplo para extender la decisión primaria basándose en las pasadas decisiones primarias para formar la decisión final, vad_flag 115. La razón de utilizar el tiempo de espera es principalmente para reducir/eliminar el riesgo de conversación a medias y el recorte al final de las rachas de voz. Sin embargo, el tiempo de espera se puede también utilizar para evitar el recorte en los pasajes musicales.

También es posible añadir un tiempo de espera adicional con el propósito de DTX. En la figura 1 esto se ha ilustrado por medio de la salida opcional vad_flag_dtx 117. Debe observarse que no es normal que exista solamente una salida vad_flag sino que la lógica del tiempo de espera utiliza otras configuraciones cuando se tiene que utilizar la salida para DTX. En esta descripción, las dos salidas de la decisión final vad_flag 115 y vad_flag_dtx 117 estarán separadas en la mayoría de las realizaciones, con el fin de simplificar la descripción. Sin embargo, se pueden aplicar otras soluciones basadas en configuraciones alternativas de tiempo de espera y de una única salida.

Existen dos razones principales para utilizar diferentes salidas de la decisión final o diferentes configuraciones de tiempos de espera dependiendo de si se va a utilizar o no la decisión VAD para DTX. En primer lugar, desde un punto de vista de la calidad de la voz existen mayores requisitos en la detección VAD cuando se utiliza para DTX. Por ello, es deseable tener la seguridad de que la conversación ha finalizado antes de conmutar al ruido de confort. La segunda motivación es que el tiempo de espera adicional se puede utilizar para la estimación de las características de ruido ambiente. Por ejemplo, en AMR NB se hace la primera estimación del ruido de confort en el descodificador basándose en el tiempo de espera específico DTX utilizado.

Como se mencionó anteriormente, existen un número de diferentes características que se pueden utilizar para la detección VAD. Una posible característica es considerar sólo la energía de la trama y compararla con un umbral para decidir si la trama contiene voz o no. Este esquema trabaja razonablemente bien para condiciones en las que la Relación Señal/Ruido (SNR) es buena pero no para los casos de baja SNR. En casos de baja SNR se utilizan preferiblemente otros sistemas, por ejemplo, comparando las características de las señales de voz y de ruido. Para ejecuciones en tiempo real, un requisito adicional en la funcionalidad VAD es la complejidad computacional, que se refleja en la frecuente representación de subbandas SNR VADs en codecs normalizados. La subbanda VAD combina normalmente las SNRs de las diferentes subbandas en un sistema común que se compara con un umbral para la decisión primaria.

La VAD 100 comprende un extractor de características 106 que proporciona la energía de la característica subbanda, y un estimador ambiental 105, que proporciona las estimaciones de la energía subbanda. Para cada trama, la VAD 100 calcula las características. Para identificar las tramas activas, la(s) característica(s) para la trama actual se compara(n) con una estimación de cómo la característica "considera" la señal ambiental.

El bloque de adición del tiempo de espera 102 se usa para extender la decisión VAD desde la VAD primaria basándose en las pasadas decisiones primarias para formar la decisión final VAD, "vad_flag", es decir las antiguas decisiones VAD se tienen también en cuenta. Como se mencionó anteriormente, la razón para utilizar el tiempo de espera es principalmente para reducir/eliminar el riesgo de conversación a medias y el recorte al final de las rachas de conversación. Sin embargo, se puede utilizar también el tiempo de espera para evitar el recorte en los pasajes musicales. Un controlador del funcionamiento 107 puede ajustar el(los) umbral(es) para el detector primario y la longitud de la adición del tiempo de espera de acuerdo con las características de la señal de entrada.

También existen soluciones conocidas en las que se utilizan múltiples características con diferentes particularidades para la decisión primaria. Para VADs basadas en el principio de la SNR de la subbanda, se ha mostrado que la introducción de una no linealidad en el cálculo de la SNR de la subbanda, a veces denominada umbrales significativos, puede mejorar el rendimiento VAD para condiciones con ruido no estacionario, por ejemplo, murmullos o ruido de oficina. Sin embargo, en estos casos existe típicamente una decisión primaria que se utiliza para añadir el tiempo de espera, que puede ser adaptativo a las condiciones de la señal de entrada, para formar la decisión final. También, muchas VADs disponen de un umbral de energía de entrada para la detección del silencio, por ejemplo, para niveles de entrada bastante bajos, la decisión primaria se fuerza al estado inactivo.

Un ejemplo en el que se utilizaron umbrales significativos para crear una solución doble VAD se describe en la solicitud publicada de patente Internacional WO2008/143569 A1. En este caso se utilizaron VADS dobles para mejorar la actualización del ruido ambiente y la detección de música. Sin embargo, sólo se utilizó una VAD agresiva primaria para decisión final vad_flag.

En el documento WO2008/143569 A1, se utilizó un método basado en una actividad de corta duración filtrada en paso bajo para detectar la existencia de música. Este método filtrado en paso bajo proporciona una cantidad que varía lentamente, apropiada para encontrar más o menos tipos continuos de sonido, típicos para, por ejemplo, música. Una decisión adicional vad_music se puede proporcionar entonces para la adición del tiempo de espera, haciéndola posible para tratar sonido musical de un modo particular.

Existen diferentes modos de generar múltiples decisiones VAD primarias. Lo más básico sería utilizar las mismas características de la VAD original pero obtener una segunda decisión primaria utilizando un segundo umbral. Otra opción es cambiar de VAD de acuerdo con las condiciones SNR estimadas, por ejemplo, utilizando la energía para las condiciones de alta SNR y cambiar al funcionamiento de la subbanda SNR para condiciones de SNR medias y bajas.

En la solicitud publicada de patente Internacional WO2011/049516 A1, se describe un detector de actividad vocal y el correspondiente método. El detector de actividad vocal está configurado para detectar la actividad vocal en una señal de entrada recibida. El VAD comprende una combinación de lógicas configuradas para recibir una señal procedente de un detector primario vocal del indicativo VAD de una decisión VAD primaria. La combinación de lógicas decide además al menos una señal procedente de un VAD externo indicativo de una decisión de la actividad vocal procedente de un VAD externo. Un procesador combina las decisiones de la actividad vocal indicadas en las señales recibidas para generar una decisión VAD primaria modificada. La decisión VAD modificada se envía a una unidad de adición del tiempo de espera.

Un problema que ocurre con el tiempo de espera es decidir cuándo y cuánto utilizar. Desde un punto de vista de la calidad de la conversación, la adición del tiempo de espera es básicamente positiva. Sin embargo, no es deseable añadir demasiado tiempo de espera ya que cualquier tiempo de espera adicional reducirá la eficacia de la solución DTX. Como no es deseable añadir el tiempo de espera a cada corta racha de actividad, existe usualmente un requisito de que haya un mínimo número de tramas activas procedente del detector primario vad_prim antes de considerar la adición de algún tiempo de espera para crear la decisión final vad_flag. Sin embargo, para evitar el recorte en la conversación es deseable mantener este número necesario de tramas activas tan bajo como sea posible.

Para el ruido no estacionario, un bajo número de tramas activas necesarias permitiría que el propio ruido genere bastantes eventos largos VAD que dispararán la adición del tiempo de espera. Así, con el fin de evitar una actividad excesiva, tal solución no permite normalmente largos tiempos de espera.

Otro problema con un número necesario de tramas activas antes de añadir el tiempo de espera para un VAD altamente eficiente es su habilidad para detectar las pausas cortas dentro de una expresión. En este caso, existe una expresión que se ha detectado correctamente pero el altavoz realiza una ligera pausa antes de continuar. Esto hace que el VAD detecte la pausa y una vez más requiera un nuevo periodo de tramas primarias activas antes de que se añada absolutamente cualquier tiempo de espera. Esto puede causar molestos efectos con recortes al final de los segmentos finales de la conversación tales como expresiones finales con estallidos no vocales.

Un ejemplo adicional de una detección de actividad vocal se describe en el documento WO2011/049514 A1, en el cual se actualiza una estimación del ruido ambiental para una señal de entrada.

Resumen

Un objetivo de las realizaciones de la invención es abordar al menos una de las cuestiones señaladas anteriormente, y este objetivo se consigue por medio de los métodos y de los aparatos de acuerdo con las reivindicaciones independientes adjuntas, y por las realizaciones de acuerdo con las reivindicaciones dependientes.

De acuerdo con un aspecto de la invención, se proporciona un método para determinar una adición del tiempo de espera en un códec de voz o de audio. Para cada trama, se determina una decisión primaria de la actividad vocal y, en función de si se tiene que realizar o no una adición del tiempo de espera de la decisión primaria, se determina una decisión final de la actividad vocal. Una medición de la actividad de corta duración y una medición de la actividad de larga duración se determinan en base a un número de tramas activas en una memoria de decisiones anteriores. La medición de la actividad de corta duración y la medición de la actividad de larga duración se comparan con un umbral determinado y se crea una decisión final alternativa para ajustar la adición del tiempo de espera si se excede el umbral.

De acuerdo con otro aspecto de la invención, se proporciona un aparato para determinar una adición del tiempo de espera. El aparato comprende medios para determinar una decisión primaria de actividad vocal para cada trama de voz o de audio y un medio para determinar una decisión final de actividad vocal en base a si se va a realizar o no una adición el tiempo de espera de la decisión primaria. El aparato comprende además medios para determinar una medición de la actividad de corta duración y una medición de la actividad de larga duración en base a un número de tramas activas en una memoria de decisiones anteriores. La medición de la actividad de corta duración y la medición de la actividad de larga duración se comparan con un umbral determinado y el aparato comprende medios para crear una decisión final alternativa para ajustar la adición del tiempo de espera si se excede el umbral.

Breve descripción de los dibujos

Para una comprensión más completa de las realizaciones de ejemplo de la presente invención, se hace ahora referencia a la siguiente descripción en conexión con los dibujos que se adjuntan, en los cuales:

La figura 1 muestra un ejemplo de una VAD genérica con estimación ambiental.

La figura 2 ilustra una realización de ejemplo de una VAD de acuerdo con la invención.

La figura 3 es un diagrama de flujo que ilustra un método VAD de ejemplo de acuerdo con una realización de la invención.

La figura 4A ilustra una realización de ejemplo de una VAD de acuerdo con la invención.

La figura 4B ilustra otra realización de ejemplo de una VAD de acuerdo con la invención.

La figura 4C ilustra otra realización más de ejemplo de una VAD de acuerdo con la invención.

La figura 5 ilustra una realización adicional de ejemplo de una VAD de acuerdo con la invención.

La figura 6 muestra una realización de una VAD con tiempo de espera.

La figura 7 muestra una realización de una VAD adicional.

Descripción detallada

Se ha encontrado actualmente una forma de atenuar tales problemas que consiste en utilizar las características temporales de los métodos de detección primaria y de los métodos de decisión final. Se ha encontrado que estos sirven para ajustar el tiempo de espera adicional. Al menos una de las decisiones primarias que entran dentro de la adición del tiempo de espera y de la decisión final extraída de la adición del tiempo de espera se utiliza preferentemente para influir en la adición del tiempo de espera, y preferiblemente se utilizan ambas. La decisión primaria que entra dentro de la adición del tiempo de espera puede ser la decisión primaria original obtenida de un detector primario de voz, o puede ser una versión modificada de tal decisión primaria original. Tal modificación se puede realizar basándose en las salidas de otras VADs.

Una realización de un tipo genérico de VAD 200 que hace uso de la decisión primaria que entra dentro de la adición del tiempo de espera 202 y de la decisión final extraída de la adición del tiempo de espera 202 se ilustra en la figura 2.

Un extractor de características 206 proporciona la energía de la característica subbanda, un estimador ambiental 205 proporciona las estimaciones de la energía subbanda, un controlador del funcionamiento 207 puede ajustar el(los) umbral(es) para el detector primario y para la longitud de la adición del tiempo de espera de acuerdo con las

características de la señal de entrada, y un detector primario de voz 201 realiza la primera decisión vad_prim 213 como se describió en conexión con la figura 1.

En esta realización, el detector de actividad vocal 200 comprende además un estimador de la actividad de corta duración 203 y/o un estimador de la actividad de larga duración 204. Las características temporales se capturan utilizando las características de la actividad de corta duración de la decisión primaria vad_prim 213, y de la actividad de larga duración de la decisión final, vad_flag de 215. Estos métodos se utilizan entonces para ajustar la adición del tiempo de espera para mejorar el rendimiento VAD para su uso en DTX por medio de crear una decisión final alternativa, vad_flag_dtx 217.

Aquí, en este caso, la actividad de corta duración se mide contando el número de tramas activas en una memoria de las últimas decisiones primarias N_st vad_prim 213. De forma similar la actividad de larga duración se mide contando el número de tramas activas en la decisión final vad_flag 215 en las últimas tramas N_lt. N_lt es mayor que N_st, preferiblemente considerablemente mayor. Estos métodos se usan entonces para crear la decisión final alternativa vad_flag_dtx 217. La ventaja de utilizar estos métodos es que simplifica el ajuste del tiempo de espera así como que es más fácil añadir el tiempo de espera solamente las veces en las que la actividad es ya alta.

Una actividad alta de larga duración indica el comienzo, el punto medio o el final de una racha activa. A primera vista este método puede parecer similar a la forma utilizada corrientemente de requerir sólo un número de tramas activas consecutivas como se mencionó anteriormente. Sin embargo, la diferencia principal es que la actividad de corta duración no se anula cuando aparece una decisión de no actividad. En su lugar, existe una memoria que recuerda una trama activa hasta las N_st tramas antes de que eventualmente salga de la memoria. Una trama no activa reducirá por consiguiente solo algo la actividad media de corta duración. Para una actividad de corta duración suficientemente alta, sería seguro añadir unas pocas tramas de tiempo de espera, ya que como la actividad de corta duración ya es alta el tiempo de espera adicional tendrá sólo un pequeño efecto sobre la actividad total. Las tramas dispersas de no actividad no reducirán suficientemente la actividad de corta duración para interrumpir tal operación de tiempo de espera.

Las tramas dispersas sin actividad pueden corresponder a pausas cortas en la mitad de una expresión o pueden ser una detección falsa sin actividad, por ejemplo, causada por secuencias cortas de conversación sin voz. Al utilizar la actividad de corta duración del modo indicado anteriormente, se puede mantener la adición del tiempo de espera durante tales ocasiones.

De manera similar, una alta actividad de larga duración indica que la racha de conversación ha sido activa durante algún tiempo. Si la actividad de larga duración es alta es así porque con gran probabilidad añadir varias tramas adicionales de tiempo de espera sólo tiene un pequeño efecto sobre la actividad total.

En una realización, la actividad de corta duración y la actividad de larga duración, respectivamente, se comparan con un umbral respectivo predeterminado. Si se alcanza el respectivo umbral, se añade un número respectivo predeterminado de tramas de tiempo de espera.

Dado que la actividad de larga duración reacciona relativamente lenta dependiendo de un final real de la actividad de conversación, existe el riesgo de utilizar un gran número de tramas de tiempo de espera añadidas un tiempo relativamente largo después del final de la racha de conversación. Con este fin, es también posible utilizar una actividad corta de baja duración como una indicación del final de una racha de conversación. Podría, por consiguiente, ser deseable en una realización, limitar la cantidad de tiempo de espera adicional si la actividad de corta duración cae por debajo de un umbral predeterminado. En otras palabras, una actividad de corta duración suficientemente baja puede ignorar la adición de tramas de tiempo de espera como indicadas por una alta actividad simultánea de larga duración.

Más adelante, las realizaciones anteriores se describen en la mayoría de los casos como modificaciones de soluciones existentes en las que el aumento de la complejidad es pequeño. Sin embargo, es también posible diseñar una VAD completamente nueva que tiene que utilizar los métodos anteriores para proporcionar una decisión VAD más fiable.

En una realización ilustrada esquemáticamente en la figura 3, un método en un detector de actividad vocal para detectar la actividad vocal en una señal de entrada recibida comprende la creación 310 de una señal indicativa de una decisión VAD primaria asociada a la señal de entrada recibida, preferiblemente analizando las características de la señal de entrada recibida. Esto se determina en 320 si se tiene que realizar o no una adición del tiempo de espera de la decisión VAD primaria. En 330 se crea una señal indicativa de una decisión VAD final. Una decisión VAD final es igual a la decisión VAD primaria si se determina que no se tiene que realizar la adición de un tiempo de espera. Una decisión VAD final es igual a una decisión de actividad vocal si se determina que se tiene que realizar la adición de un tiempo de espera. Ya que se añade el tiempo de espera, la decisión de actividad vocal se fija para indicar la trama activa, es decir, una trama que contiene más conversación que ruido. Una medición de la actividad de corta duración se deduce en 340 de las últimas decisiones VAD primarias N_st y/o una medición de la actividad de larga duración se deduce en 342 en las últimas decisiones VAD finales N_lt. La determinación de si se tiene que realizar o no la adición de un tiempo de espera se hace dependiendo de la medición de la actividad de corta duración y/o de

la medición de la actividad de larga duración. Aunque si la figura 3 se ilustra como un simple flujo de eventos, el sistema real tratará una trama a continuación de la otra. Las flechas discontinuas indican que la dependencia de la medición de la actividad de corta duración y/o de la medición de la actividad de larga duración es válida para una trama sucesiva.

Se debe comprender que la figura 3 no ilustra un flujo de señal sino más bien etapas del método a realizar de acuerdo con una realización de la invención. Es decir, la creación de una decisión VAD final 330 puede comprender la creación de una decisión final alternativa (por ejemplo, vad_flag_dtx 217) basándose en las mediciones de la actividad de corta duración y/o de la actividad de larga duración. La decisión final alternativa, sin embargo, no se utiliza como una entrada para el estimador de la actividad de larga duración 204 ya que introduciría un bucle de realimentación de la actividad (debido a la modificación de la característica a medir con la adición del tiempo de espera ajustado). Por consiguiente, la creación de una decisión VAD final 330 puede también comprender crear una decisión final (por ejemplo, vad_flag 215) basándose en la técnica tradicional del tiempo de espera y/o en las mediciones de la actividad de corta duración pero no en las mediciones de la actividad de larga duración, las cuales se utilizan entonces como una entrada para el estimador de la actividad de larga duración 204, como se muestra en la figura 2.

En una realización, ilustrada esquemáticamente en la figura 4A, un detector de actividad vocal 400 comprende una sección de entrada 412, una disposición del detector primario de voz 401 y una unidad de adición del tiempo de espera 402. La sección de entrada está configurada para recibir una señal de entrada. La disposición del detector primario de voz 401 está conectada a la sección de entrada 412. La disposición del detector primario de voz 401 está configurada para detectar la actividad vocal en la señal de entrada recibida y para crear una señal indicativa de una decisión VAD primaria asociada a la señal de entrada recibida. La unidad de adición del tiempo de espera 402 está conectada a la disposición del detector primario de voz 401. La unidad de adición del tiempo de espera 402 está configurada para determinar si se tiene que realizar o no la adición de un tiempo de espera de dicha decisión VAD primaria y para crear una señal indicativa de una decisión VAD final. La decisión VAD final es igual a la decisión VAD primaria si se determina que no se debe realizar la adición de un tiempo de espera. La decisión VAD final es igual a una decisión de la actividad vocal si se determina que se debe realizar la adición del tiempo de espera. El detector de actividad vocal 400 comprende además un estimador de la actividad de corta duración 403 y/o un estimador de la actividad de larga duración 404. El estimador de actividad de corta duración 403 está conectado a una entrada de una unidad de adición del tiempo de espera 402. El estimador de la actividad de corta duración 403 está configurado para deducir una medición de la actividad de corta duración de las últimas decisiones VAD primarias N_st. El estimador de la actividad de larga duración 404 está conectado a una salida de la unidad de adición del tiempo de espera 402. El estimador de la actividad de larga duración 404 está configurado para deducir una medición de la actividad de larga duración de las últimas decisiones VAD finales N_lt. La unidad de adición del tiempo de espera 402 está conectada a una salida del estimador de la actividad de corta duración 403 y/o del estimador de la actividad de larga duración 404. La unidad de adición del tiempo de espera 402 está además configurada para realizar la determinación del tiempo de espera dependiendo de la medición de la actividad de corta duración y/o de la medición de la actividad de larga duración. La determinación del tiempo de espera dependiente de la medición de la actividad de corta duración y/o de la medición de la actividad de larga duración puede entonces utilizarse para ajustar la adición del tiempo de espera para mejorar el rendimiento VAD para su uso en DTX al crear una decisión final alternativa.

El detector de actividad vocal se proporciona normalmente en un códec vocal o de sonido. Tales codecs se proporcionan típicamente en diferentes dispositivos finales, por ejemplo, en redes de telecomunicación. Ejemplos no limitativos son los teléfonos, ordenadores, etc., en los que se realiza la detección o registros del sonido.

En una realización, la decisión VAD final la da un indicador adicional 410, además de la decisión VAD final hecha sin la utilización de las mediciones de la actividad de corta duración o de las mediciones de la actividad de larga duración, normalmente como una decisión VAD final para el uso en DTX, como se ilustra en la figura 4B. Las dos versiones de las decisiones finales se pueden utilizar entonces en paralelo para diferentes unidades o funcionalidades. En otra realización alternativa, la utilización de las mediciones de la actividad de corta duración o de las mediciones de la actividad de larga duración se puede conmutar si/no dependiendo del contexto en el cual se va a utilizar la decisión VAD.

En otra realización, en la que una decisión VAD final no está disponible o no es apropiada para hacer cualquier análisis de la actividad de larga duración, se puede realizar en su lugar un análisis de la actividad de larga duración sobre la decisión VAD primaria. En tal realización, el estimador de la actividad de larga duración 404 se conecta por el contrario a la entrada de la unidad de adición del tiempo de espera 402, como se muestra en la figura 4C, y se deduce una medición de la actividad de larga duración de las últimas decisiones VAD primarias N_lt.

En otra realización más, se pueden realizar las estimaciones de la actividad de corta y larga duración en la decisión VAD primaria y/o final diferente de la decisión VAD primaria y/o final sobre la cual se tiene que realizar el ajuste de la adición del tiempo de espera. Una posibilidad es tener una simple VAD que produzca una decisión VAD primaria y una simple unidad de tiempo de espera modificando la dentro de una decisión VAD final. El comportamiento de la actividad de corta y de larga duración de tales decisiones VAD primarias y/o finales puede entonces ser analizado. Sin embargo, otra configuración VAD, por ejemplo una más sofisticada, se puede entonces utilizar para proporcionar

la decisión VAD primaria de interés para el ajuste de la adición del tiempo de espera. Las actividades analizadas a partir del sistema simple pueden entonces utilizarse para controlar el funcionamiento de la unidad de adición del tiempo de espera 402 del sistema VAD más elaborado, proporcionando una decisión VAD final fiable.

A partir aquí, se describirá un ejemplo de una realización de un detector de actividad vocal 500 con referencia a la figura 5. Esta realización se basa en un procesador 510, por ejemplo, un microprocesador, que ejecuta un componente de software 501 para crear una señal indicativa de una decisión VAD primaria, un componente de software 502 para determinar si se tiene que realizar una adición del tiempo de espera de la decisión VAD primaria y un componente de software 503 para crear una señal indicativa de una decisión VAD final. En esta realización el procesador 510 ejecuta un componente de software 504 para deducir la medición de una actividad de corta duración procedente de las últimas decisiones VAD primarias N_st y/o un componente de software 505 para deducir la medición de una actividad de larga duración procedente de las últimas decisiones VAD finales N_lt. Estos componentes de software se almacenan en una memoria 520. El procesador 510 se comunica con la memoria 520 sobre un bus del sistema 515. La señal de audio es recibida por un controlador de entrada/salida (E/S) 530 que controla un bus E/S 516, al cual están conectados el procesador 510 y la memoria 520. En esta realización, las señales recibidas por el controlador E/S 530 se almacenan en la memoria 520, en la cual son tratadas por los componentes de software. El componente de software 501 puede realizar la funcionalidad de la etapa 310 en la realización descrita con referencia a la figura 3 anterior.

El componente de software 502 puede realizar la funcionalidad de la etapa 320 en la realización descrita con referencia a la figura 3 anterior. El componente de software 503 puede realizar la funcionalidad de la etapa 330 en la realización descrita con referencia a la figura 3 anterior. El componente de software 504 puede realizar la funcionalidad de la etapa 340 en la realización descrita con referencia a la figura 3 anterior. El componente de software 505 puede realizar la funcionalidad de la etapa 342 en la realización descrita con referencia a la figura 3 anterior.

La unidad E/S 530 se puede interconectar al procesador 510 y/o a la memoria 520 por medio de un bus E/S 516 para habilitar la entrada y/o la salida de datos relevantes tales como las señales de entrada y las decisiones VAD finales.

En una realización, los contadores de tramas activas en la memoria de las decisiones primarias y de las decisiones finales se utilizan como se ha descrito anteriormente. En realizaciones alternativas, sería posible utilizar una ponderación que dependa de la antigüedad de la trama activa en la memoria. Esto es posible tanto para la actividad primaria de corta duración como para la actividad de decisión final de larga actividad. En realizaciones adicionales, podría ser posible utilizar diferentes tiempos de espera adicionales dependiendo de otras características de la señal de entrada, tales como nivel de conversación, nivel de ruido y/o SNR estimados.

En otras realizaciones, podría ser interesante utilizar más de dos características temporales para localizar mejor el principio, la mitad o el final de una racha activa de conversación.

En realizaciones adicionales, los principios de decisiones de los tiempos de espera descritos anteriormente se podrían también combinar con otras soluciones de mejora de la VAD tales como los principios del combinador Multi VAD presentado en el documento WO2011/049516. En este caso se puede utilizar la decisión VAD primaria modificada como entrada al estimador de la actividad de corta duración y el bloque de adición del tiempo de espera. El combinador Multi VAD podría entonces ser considerado como una parte de la disposición del detector vocal primario.

De manera similar, se pueden integrar con las presentes ideas, ventajosa y fácilmente, diferentes enfoques adicionales para estimar el ambiente.

Un códec G.718 de acuerdo con las normas 3GPP2 se utiliza como la base para una realización que se presentará en este documento más adelante. Una descripción detallada de las partes relacionadas se puede encontrar en, por ejemplo, la solicitud publicada de patente internacional WO2009/000073 A1.

La figura 6 muestra un diagrama de bloques de un sistema de comunicación de sonidos del documento WO2009/000073 A1 que comprende un preprocesador 601, un analizador de espectros 602, un detector de actividad del sonido 603, un estimador del ruido 604, un reductor opcional del ruido 605, un analizador LP y localizador del tono 606, un módulo de actualización de la energía estimada del ruido 607, un clasificador de la señal 608 y un codificador de sonido 609. La detección de la actividad del sonido (primera fase de la clasificación de la señal) se realiza en el detector de la actividad del sonido 603 utilizando las estimaciones de la energía de ruido calculadas en la trama anterior. La salida del detector de actividad del sonido 603 es una variable binaria utilizada posteriormente por el codificador 609 y que determina si la trama actual se codifica como activa o como inactiva.

El módulo “SAD Basado en SNR” 603 es el módulo en el que se pueden practicar las realizaciones de la presente descripción. Actualmente, la realización presentada sólo cubre la cadena de señal de banda ancha, con muestreo en 16kHz, pero una modificación similar podría ser beneficiosa para la cadena de señal de banda estrecha, con muestreo en 8 kHz, en cualesquiera otras tasas de muestreo.

En una realización, basada en los principios presentados en el documento WO2011/049516 A1, se utiliza la VAD original según el documento WO2009/000073 A1 (VAD 1) como la primera VAD, generando las señales localVAD y vad_flag. Esta localVAD se usa en la presente descripción como VAD_prim 213 en la cual se hace la estimación de la actividad de corta duración.

La VAD adicional (VAD 2) se basa también en el documento WO2009/000073 A1 pero se logra utilizando las modificaciones para la estimación de ruido ambiental y para la SAD basada en SNR. La figura 7 muestra un diagrama de bloques para la segunda VAD. El diagrama de bloques muestra un preprocesador 701, un analizador de espectro 702, un módulo “SAD basado en SNR” 703, un estimador del ruido 704, un reductor opcional del ruido 705, un analizador de LP y localizador del tono 706, un módulo de actualización de la energía estimada de ruido 707, un clasificador de señal 708 y un codificador de sonido 709.

El diagrama de bloques también muestra las decisiones VAD primarias y final para VAD 2, localVAD_he 710 y vad_flag_he 711, respectivamente. La localVAD_he 710 y la vad_flag_he 711 se utilizan en el detector primario de voz de la VAD1 para producir la localVAD.

Para esta realización se añaden las variables siguientes al estado del codificador (Encoder_State):

```
long long vad_flag_reg;        /* memory of old vad_flag */
long long vad_prim_reg;        /* memory of old localVAD */
short vad_flag_cnt_50;         /* counter of vad_flag active frames */
short vad_prim_cnt_16;         /* counter of primary active frames */

short hangover_cnt_dtx;       /* counter of hangover frames for DTX */
```

Todos estos estados se deben poner a cero durante la inicialización, es decir, se podría hacer en la rutina wb_vad_init().

Además, se actualizan las características de la actividad de corta duración y de la actividad de larga duración, lo cual se debe hacer al final del tratamiento para cada trama. Se puede hacer añadiendo el siguiente código en el fichero fuente apropiado:

```
if ((st->vad_flag_reg & (long long) 0x01LL << 49) != 0)
{
   st->vad_flag_cnt_50=st->vad_flag_cnt_50-1;
}
st->vad_flag_reg = (st->vad_flag_reg & (long long)
0x3fffffffffffffLL) << 1;
if (vad_flag)
{
   st->vad_flag_reg = st->vad_flag_reg | 0x01L;
   st->vad_flag_cnt_50 = st->vad_flag_cnt_50+1;
}
if ((st->vad_prim_reg & (long long) 1LL << 15) != 0)
{
   st->vad_prim_cnt_16=st->vad_prim_cnt_16-1;
}
st->vad_prim_reg = (st->vad_prim_reg & (long long)
0x3fffffffffffffLL) << 1;
if (localVAD)
{
   st->vad_prim_reg = st->vad_prim_reg | 0x01L;
   st->vad_prim_cnt_16 = st->vad_prim_cnt_16+1;
}
```

Aquí la variable st se refiere a la variable Encoder_State asignada al codificador. Así, para la trama siguiente, las variables de estado st->vad_flag_cnt_50 contendrán la actividad de la decisión final de larga duración en la forma del número de tramas que son activas dentro de las últimas 50 tramas y la variable del estado st->vad_prim_cnt_16 contendrá la actividad primaria de corta duración en la forma del número de tramas primarias activas dentro de las últimas 16 tramas. La longitud de la memoria de la actividad de corta duración, 16 tramas, y la longitud de la

memoria de la actividad de larga duración, 50 tramas, son valores que se usan en esta realización en particular. Estas cifras son valores típicos que se pueden utilizar en una ejecución operativa, pero los valores absolutos no son cruciales. Estos números pueden por consiguiente adaptarse según diferentes tipos de ejecuciones, por ejemplo, como un ajuste de las propiedades de los tiempos de espera. Generalmente, la longitud de la memoria de la actividad de larga duración es mayor que la longitud de la memoria de la actividad de corta duración, y preferiblemente considerablemente mayor, como en el ejemplo presentado anteriormente. En una realización típica, la relación entre la longitud de la memoria de la actividad de larga duración y la longitud de la memoria de la actividad de corta duración está dentro del intervalo de 2,5 a 5. Esta relación también se puede adaptar a diferentes tipos de ejecuciones en las cuales se espera que se presenten frecuentemente diferentes tipos de sonido.

El código para decidir cuánto tiempo de espera, hangover_short, se debe añadir, se puede realizar utilizando la siguiente modificación del código en la cual:

lp_snr

es una estimación de la SNR filtrada en paso bajo

th_clean

utiliza el umbral de la SNR para decidir si la entrada está libre de conversación

thr1

el umbral calculado para el detector primario

```
if (lp_snr < th_clean)
{
   thr1 = nk * lp_snr + nc; /* Linear function for noisy speech */
   if (st->Opt_SC_VBR)
   {
       hangover_short = 1;
   }
   else
   {
       hangover_short = 4;
   }
}
else
{
   thr1 = sk * lp_snr + sc; /* Linear function for clean speech */
   hangover_short = 1;
}
```

A lo siguiente que añade entonces el código necesario para la adaptación del tiempo de espera utilizado para DTX hangover_short_dtx.

```
if (lp_snr < th_clean)
{
   thr1 = nk * lp_snr + nc; /* Linear function for noisy speech */
   if (st->Opt_SC_VBR)
   {
       hangover_short = 1;
   }
   else
   {
       hangover_short = 4;
   }
}
else
{
   thr1 = sk * lp_snr + sc; /* Linear function for clean speech */
   hangover_short = 1;
}
hangover_short_dtx = hangover_short; /* start with same hangover for
DTX */
if (st->Opt_DTX_ON)
{
   if (st->vad_prim_cnt_16 > 12) /* 12 requires roughtly > 80%
primary activity */
   {
       hangover_short_dtx = hangover_short_dtx + 1;
   }
   if (st->vad_flag_cnt_50 > 40) /* 40 requires roughtly > 80% flag

activity */
   {
       hangover_short_dtx = hangover_short_dtx + 3;
   }
   /* Keep hangover_short lower than maximum hangover count */
   if (hangover_short_dtx > HANGOVER_LONG-1)
   {
       hangover_short_dtx=HANGOVER_LONG-1;
   }
   /* Only allow short HO if not sufficient active frames */
   if (st->vad_prim_cnt_16 < 7 && hangover_short_dtx > 4)
   {
       hangover_short_dtx=4;
   }
}
```

Aquí también, existe un número de figuras especificadas, que se tienen que considerar como variables del diseño. Estos números pueden por consiguiente también ser adaptados en diferentes tipos de realizaciones, por ejemplo, como un ajuste de las propiedades del tiempo de espera.

El código para realizar el tiempo de espera real se puede hacer con la siguiente modificación:

flag	La decisión VAD final incluyendo el tiempo de espera
localVAD	Decisión primaria
snr_sum	Característica VAD en la forma de estimación de una SNR subbanda
st->nb_active_frames	Número de tramas activas consecutivas (decisiones primarias)

st->hangover_cnt Contador de las tramas de tiempo de espera utilizadas

```
flag = 0;
*localVAD = 0;
if (snr_sum > thr1 && (st->Opt_HE_SAD_ON == 0 || (flag_he == 1 &&
flag_he1 == 1))) /* Speech present */
{
   flag = 1;
   if (snr_sum > thr1)
   {
       *localVAD = 1; /* VAD without hangover */
   }
   st->nb_active_frames++; /* Counter of consecutive active speech
frames */
   if (st->nb_active_frames >= ACTIVE_FRAMES)
   {
       st->nb_active_frames = ACTIVE_FRAMES;
       st->hangover_cnt = 0; /* Reset the counter of hangover
frames after at least "active_frames" speech frames */
   }
   /* inside HO period */
   if (st->hangover_cnt < HANGOVER_LONG && st->hangover_cnt != 0)
   {
       st->hangover_cnt++;
   }
}
else
{ /* Reset the counter of speech frames necessary to start hangover
algorithm */
   st->nb_active_frames = 0;
   if (st->hangover_cnt < HANGOVER_LONG) /* inside HO period */
   {
       st->hangover_cnt++;
   }
   if (st->hangover_cnt <= hangover_short) /* "hard" hangover */
   {
       flag = 1 ;
   }
```

Esto se modifica en lo que sigue para incluir la nueva decisión VAD a utilizar para DTX, vad_flag_dtx. Utilizando la adaptación del tiempo de espera DTX definida anteriormente, hangover_short_dtx. La cual añade las siguientes variables:

flag_dtx Decisión VAD final que incluye también el tiempo de espera específico DTX

st->hangover_cnt_dtx Contador del número de tramas de tiempo de espera utilizadas para DTX

```
flag = 0;
flag_dtx = 0;
*localVAD = 0;
if (snr_sum > thr1 && (st->Opt_HE_SAD_ON == 0 || (flag_he == 1 &&
flag_he1 == 1))) /* Speech present */
{
   flag = 1;
   flag_dtx=1;
   if (snr_sum > thr1)
   {
       *localVAD = 1; /* VAD without hangover */
   }
   st->nb_active_frames++; /* Counter of consecutive active speech
frames */
   if (st->nb_active_frames >= ACTIVE_FRAMES)
   {
      st->nb_active_frames = ACTIVE_FRAMES;
      st->hangover_cnt = 0; /* Reset the counter of hangover frames
after at least "active_frames" speech frames */
   }
   if (st->Opt_DTX_ON)
   {
       if (st->vad_flag_cnt_50 > 45) /* 45 requires roughtly > 90%
flag activity */
       {
          /* If sufficient activity during last second add hangover
             with out requirement for active frames
          */
          st->hangover_cnt_dtx=0;
       }
   }
   /* inside HO period */
   if (st->hangover_cnt < HANGOVER_LONG && st->hangover_cnt != 0)
   {
       st->hangover_cnt++;
   }
   if (st->hangover_cnt_dtx < HANGOVER_LONG && st->hangover_cnt_dtx
!= 0)
   {
       st->hangover_cnt_dtx++;
   }
}
else
{ /* Reset the counter of speech frames necessary to start hangover
algorithm */
   st->nb_active_frames = 0;
   if (st->hangover_cnt < HANGOVER_LONG) /* inside HO period */
   {
       st->hangover_cnt++;
   }
   if (st->hangover_cnt <= hangover_short) /* "hard" hangover */
   {
       flag = 1 ;
       flag_dtx = 1 ;
   }
   if (st->hangover_cnt_dtx < HANGOVER_LONG) /* inside HO period
```

```
*/
    {
        st->hangover_cnt_dtx++;
    }
    if (st->hangover_cnt_dtx <= hangover_short_dtx) /* "hard"
hangover */
    {
        flag_dtx = 1;
    }
```

Con la utilización de las características de la actividad de corta duración de la decisión primaria y de la actividad de larga duración de la decisión final es posible añadir un tiempo de espera extra más específicamente dentro de la rachas de conversación y al final de la racha, reduciendo por consiguiente la cantidad de recortes de la conversación, en particular para VADs de alto rendimiento.

La actividad de larga duración de la decisión final también hace posible añadir tiempo de espera a las rachas cortas después de largas expresiones, lo cual reduce el riesgo de recorte trasero final de los estallidos sin voz.

Con el uso de las características de la actividad, se hace posible aumentar el tiempo de espera en segmentos con la ya alta actividad de conversación. Esto permite alargar la extensión sin el riesgo de aumentar dramáticamente la actividad total.

Con las características adicionales, como se ha presentado además anteriormente, es posible un refinamiento adicional que hace posible la extensión del tiempo de espera incluso en las condiciones más limitadas, tales como un bajo nivel de conversación.

Con una SAD más agresiva podría ser más sencillo eliminar cualquier recorte de la conversación añadiendo algún tiempo de espera extendido, particularmente si se pueda hacer más específicamente para ya los elementos de alta actividad. Esta solución podría ser más sencilla de ajustar que tratar de reajustar una solución basada en diversas SADs funcionando en paralelo.

Las realizaciones descritas anteriormente deben ser comprendidas como unos pocos ejemplos ilustrativos de las ideas presentes. Los expertos en la técnica entenderán que se pueden hacer diversas modificaciones, combinaciones y cambios en las realizaciones sin apartarse del ámbito general de la presente realización. En particular, se pueden combinar diferentes soluciones parciales en las diferentes realizaciones con otras configuraciones, en aquello que sea técnicamente posible.

REIVINDICACIONES

1. Un método para determinar una adición del tiempo de espera para un códec de voz o audio, donde para cada trama se determina una decisión primaria de la actividad vocal y, dependiendo de si se debe realizar una adición del tiempo de espera de la decisión primaria o no, se determina una decisión final de la actividad de vocal, comprendiendo el método:

- determinar una medición de la actividad de corta duración en base a un número de tramas activas en una memoria de las últimas decisiones primarias N_st;

- determinar una medición de la actividad de larga duración en base a un número de tramas activas en una memoria de las últimas decisiones finales N_lt;

- comparar la medición de la actividad de corta duración con un primer umbral y la medición de la actividad de larga duración con un segundo umbral;

- crear una decisión final alternativa para ajustar la adición del tiempo de espera mediante un número predeterminado de tramas de tiempos de espera si se excede al menos uno de los umbrales primero y segundo.

2. El método de acuerdo con la reivindicación 1, en el que N_lt es mayor que N_st.

3. El método de acuerdo con la reivindicación 1 o 2, en el que N_st es 16 y N_lt es 50.

4. El método de acuerdo con cualquiera de las reivindicaciones 1 a 3, en el que el primer umbral es 12 y el segundo umbral es 40.

5. El método de acuerdo con cualquiera de las reivindicaciones 1 a 4, en el que la decisión final alternativa se determina para uso en transmisión discontinua (DTX).

6. El método de acuerdo con cualquiera de las reivindicaciones 1 a 5, en el que la decisión final alternativa corresponde a vad_flag_dtx.

7. Un aparato para determinar una adición del tiempo de espera, comprendiendo el aparato:

- medios para determinar una decisión primaria de actividad vocal para cada trama de voz o audio;

- medios para determinar una decisión final de la actividad vocal en función de si se debe realizar o no la adición del tiempo de espera de la decisión primaria;

- medios para determinar una medición de la actividad de corta duración en base a un número de tramas activas en una memoria de las últimas decisiones primarias N_st;

- medios para determinar una medición de la actividad de larfa duración en base a un número de tramas activas en una memoria de las últimas decisiones N_lt finales;

- medios para comparar la medición de la actividad de corta duración con un primer umbral y la medición de la actividad de larga duración con un segundo umbral;

- medios para crear una decisión final alternativa para ajustar la adición del tiempo de espera mediante un número predeterminado de tramas de tiempo de espera si se excede al menos uno de los umbrales primero y segundo.

8. El aparato de acuerdo con la reivindicación 7, en el que N_lt es mayor que N_st.

9. El aparato de acuerdo con la reivindicación 7 u 8, en el que N_st es 16 y N_lt es 50.

10. El aparato de acuerdo con cualquiera de las reivindicaciones 7 a 9, en el que el primer umbral es 12 y el segundo umbral es 40.

11. El aparato de acuerdo con cualquiera de las reivindicaciones 7 a 10, en el que la decisión final alternativa se determina para uso en transmisión discontinua (DTX).

12. El aparato de acuerdo con cualquiera de las reivindicaciones 7 a 11, en el que la decisión final alternativa corresponde a vad_flag_dtx.

13. El aparato de acuerdo con cualquiera de las reivindicaciones 7 a 12, en el que el aparato está comprendido en un códec de voz o de audio.

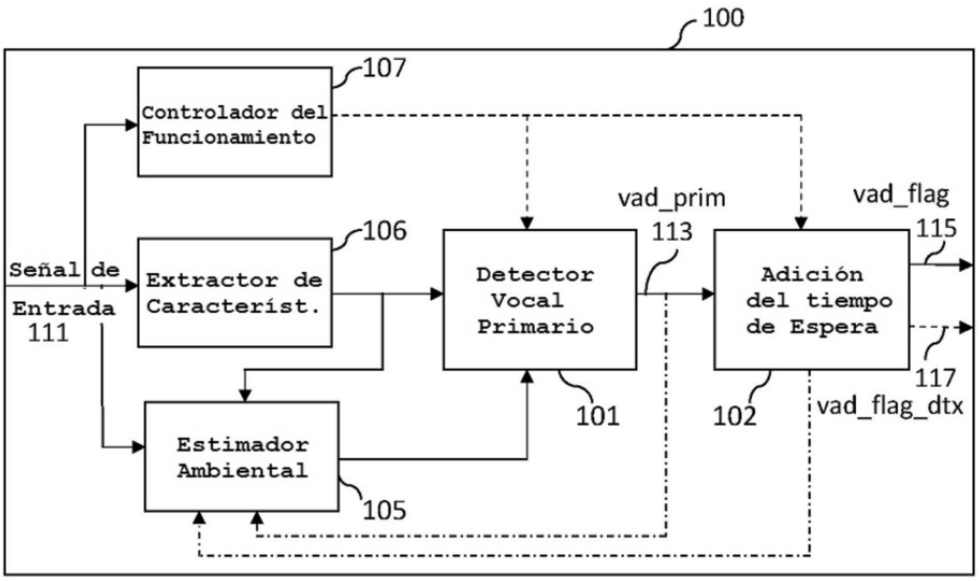

FIGURA 1

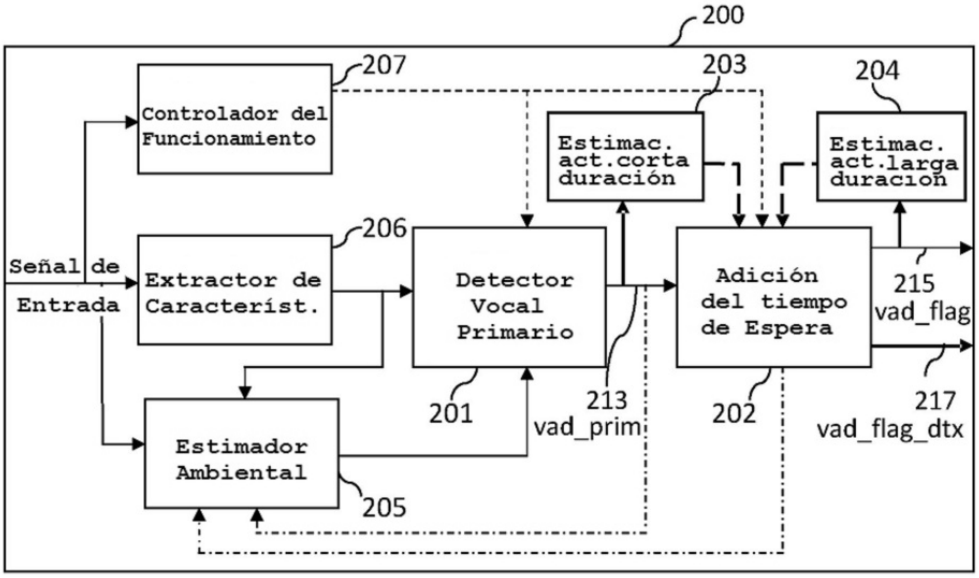

FIGURA 2

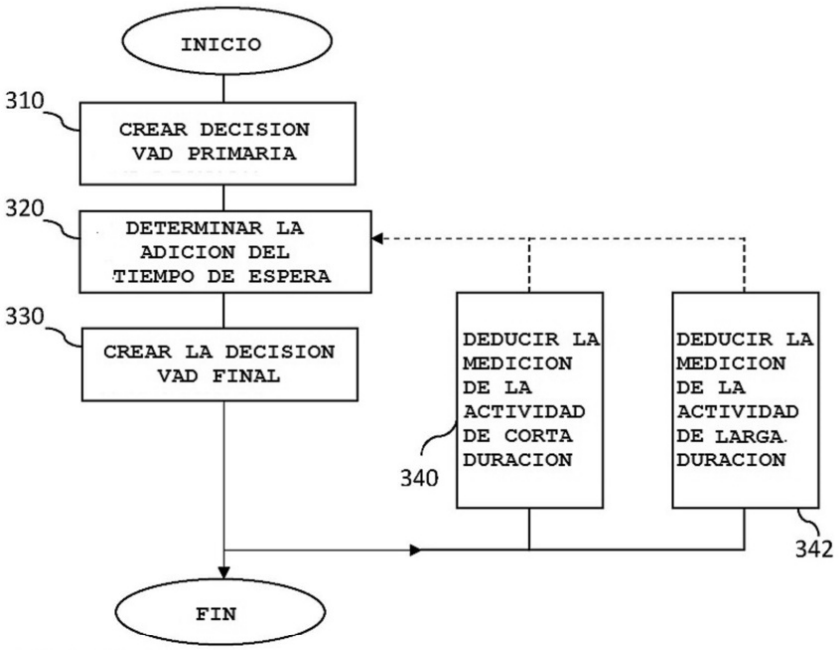

FIGURA 3

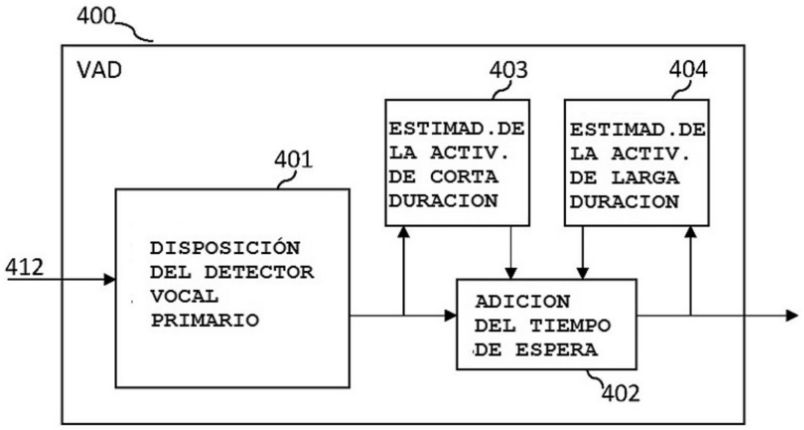

FIGURA 4A

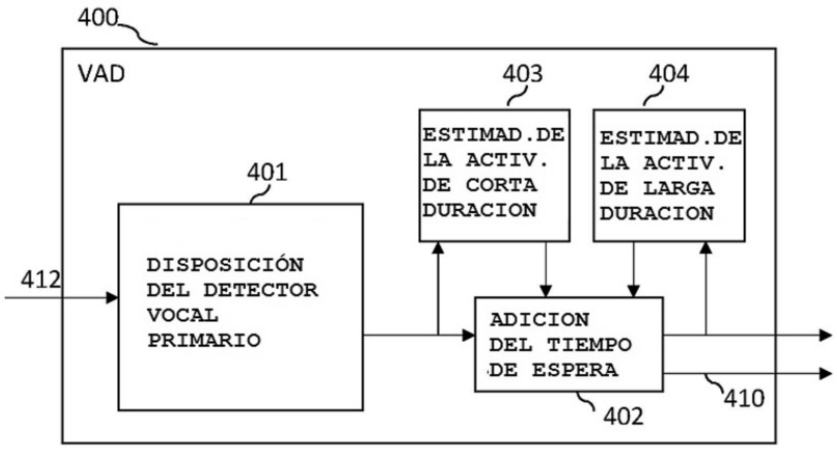

FIGURA 4B

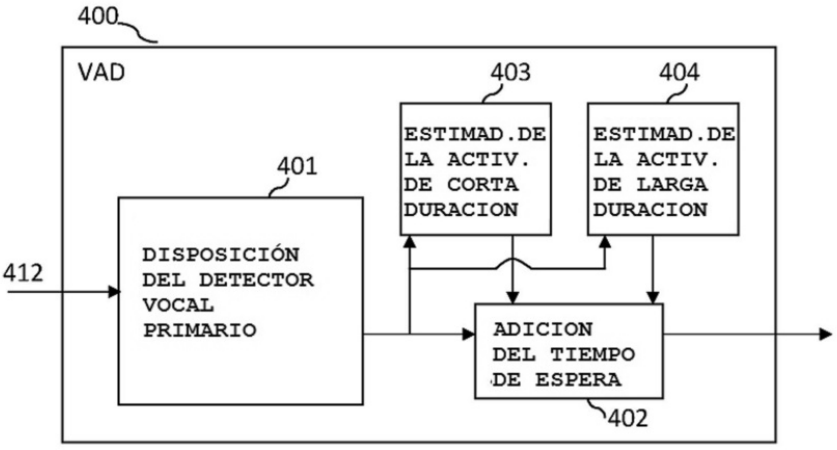

FIGURA 4C

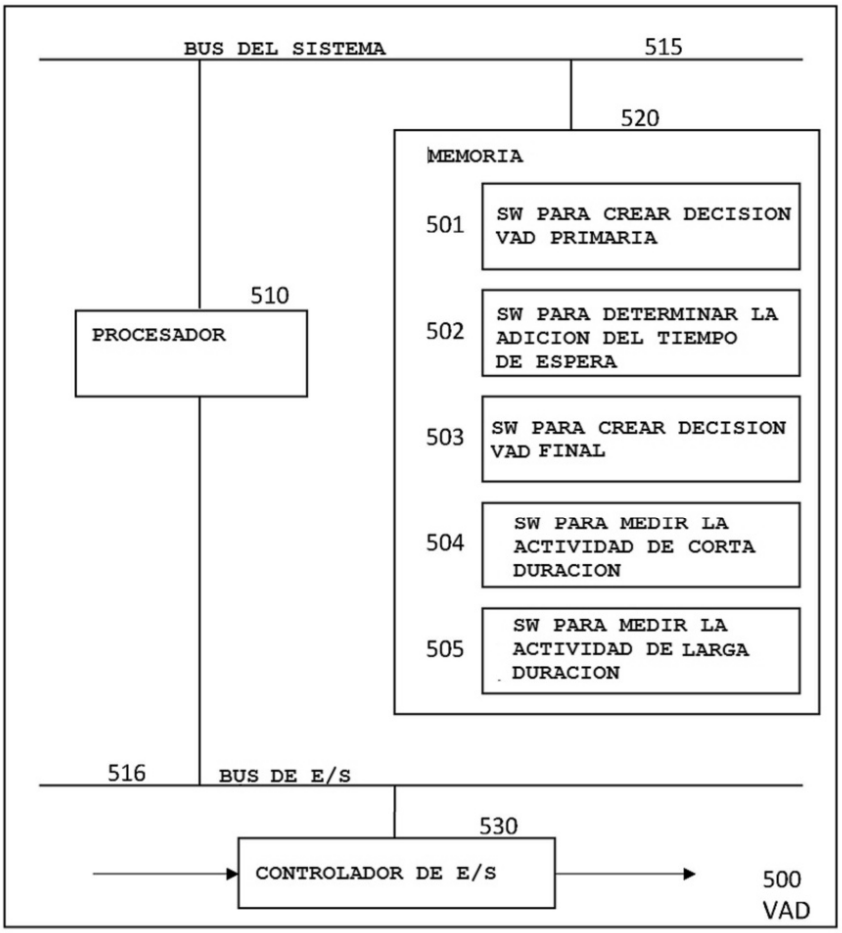
BUS DEL SISTEMA
515
520
MEMORIA
501
SW PARA CREAR DECISION VAD PRIMARIA
502
SW PARA DETERMINAR LA ADICION DEL TIEMPO DE ESPERA
503
SW PARA CREAR DECISION VAD FINAL
504
SW PARA MEDIR LA ACTIVIDAD DE CORTA DURACION
505
SW PARA MEDIR LA ACTIVIDAD DE LARGA DURACION
510
PROCESADOR
516
BUS DE E/S
530
CONTROLADOR DE E/S
500
VAD

FIGURA 5

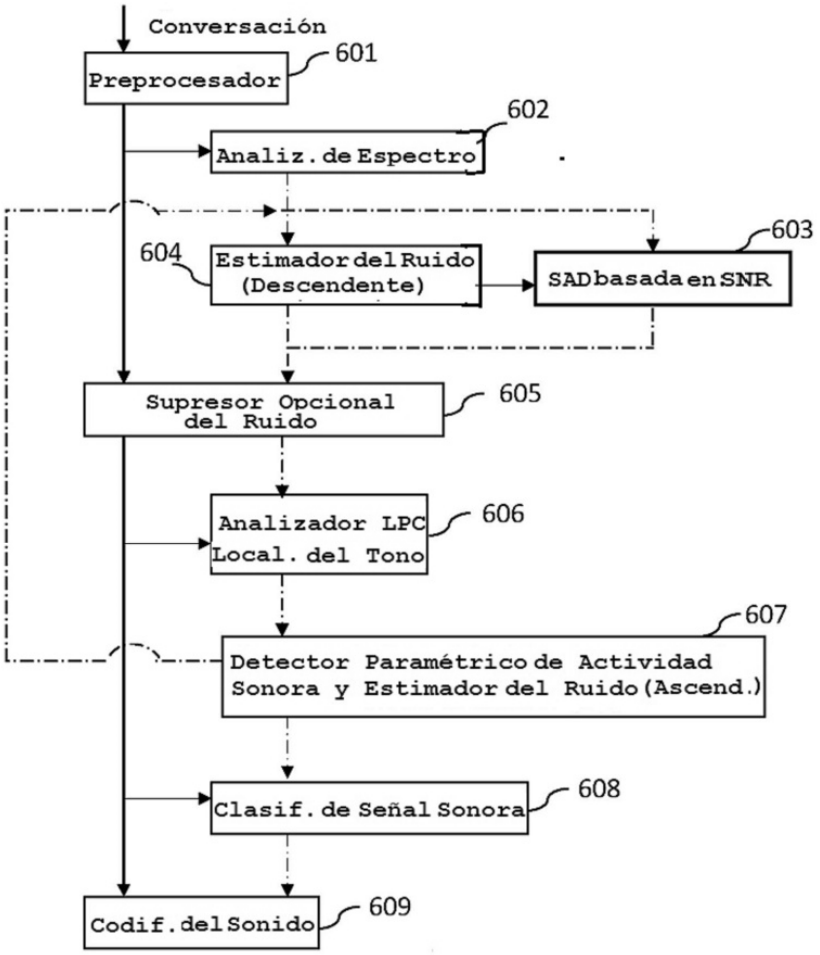
Conversación
Preprocesador
601
Analiz. de Espectro
602
Estimador del Ruido (Descendente)
604
SAD basada en SNR
603
Supresor Opcional del Ruido
605
Analizador LPC Local. del Tono
606
Detector Paramétrico de Actividad Sonora y Estimador del Ruido (Ascend.)
607
Clasif. de Señal Sonora
608
Codif. del Sonido
609

FIGURA 6

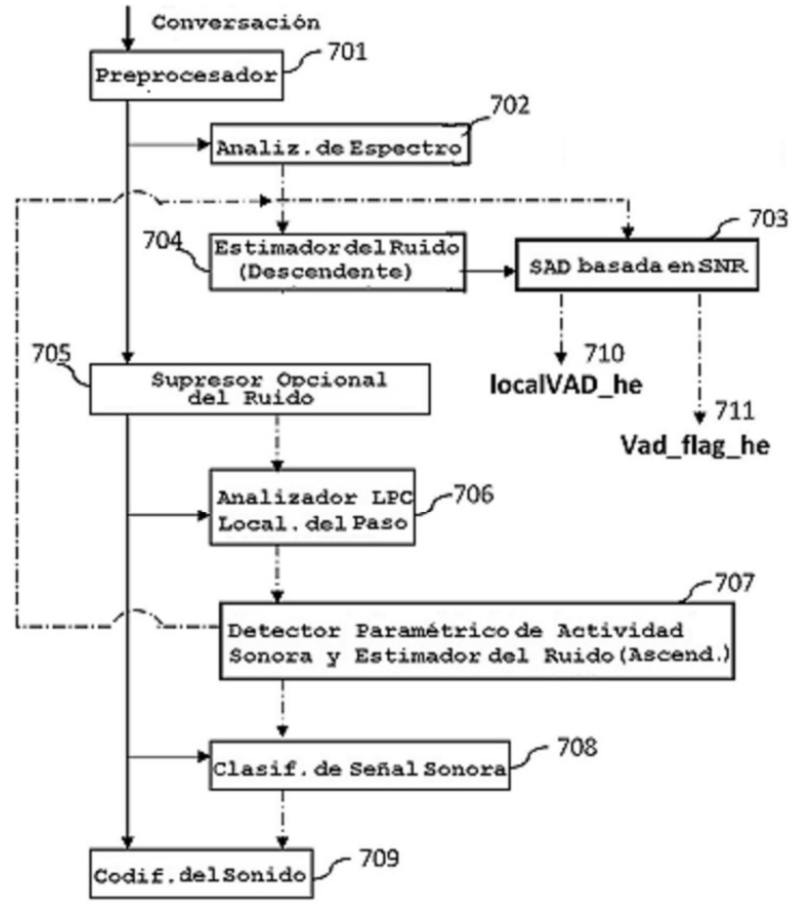
Conversación
Preprocesador
701
Analiz. de Espectro
702
Estimador del Ruido (Descendente)
704
SAD basada en SNR
703
710
localVAD_he
711
Vad_flag_he
Supresor Opcional del Ruido
705
Analizador LPC Local. del Paso
706
Detector Paramétrico de Actividad Sonora y Estimador del Ruido (Ascend.)
707
Clasif. de Señal Sonora
708
Codif. del Sonido
709

FIGURA 7