(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 525 427**

(51) Int. CI.:

***G10L 25/78*** (2013.01)
***G10L 19/02*** (2013.01)
***G10L 21/0208*** (2013.01)
*G10L 21/0232* (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **09.02.2007 E 07709334 (2)**

(97) Fecha y número de publicación de la concesión europea: **24.09.2014 EP 1982324**

(54) Título: **Un detector de voz y un método para suprimir sub-bandas en un detector de voz**

(30) Prioridad:

**10.02.2006 US 743276 P**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
**22.12.2014**

(73) Titular/es:

**TELEFONAKTIEBOLAGET L M ERICSSON (PUBL) (100.0%)**
**164 83 Stockholm, SE**

(72) Inventor/es:

**SEHLSTEDT, MARTIN**

(74) Agente/Representante:

**DE ELZABURU MÁRQUEZ, Alberto**

## DESCRIPCIÓN

Un detector de voz y un método para suprimir sub-bandas en un detector de voz

### Campo técnico

La presente invención está relacionada con un detector de voz, un detector de actividad de la voz (VAD) y un método para suprimir selectivamente las sub-bandas en un detector de voz.

### Antecedentes

Una parte importante para reducir la tasa de bits en codificadores del habla de alto rendimiento es el uso del ruido de confort en lugar del silencio o rebajar la tasa de bits de fondo. La función clave que hace posible esto es un detector de actividad de la voz (VAD), que permite la separación entre el habla y el ruido de fondo.

Se han propuesto diversos tipos de detectores de actividad de voz, y en la TS 26.094, véase la referencia [1] se divulga un VAD (aquí denominado AMR VAD 1) y variantes en la referencia [3]. Las características básicas del AMR VAD 1 son:

- detector de la suma de la relación señal-ruido (SNR) de la sub-banda,

- adaptación del umbral basándose en el nivel de la señal,

- adaptación de la estimación del fondo basándose en decisiones previas, y

- análisis de recuperación del estancamiento para aumentos escalonados del nivel de ruido.

Un inconveniente del AMR VAD 1 es que es extra-sensible para algunos tipos de ruido de fondo no estacionario.

Otro VAD (denominado aquí EVRC VAD) se divulga en la C.s0014-A, ver referencia [2], como EVRC RDA y la referencia [4]. Las principales tecnologías utilizadas son:

- análisis de banda repartida, donde la banda del caso peor se utiliza para la selección de velocidad en un códec de habla de velocidad variable.

- se utiliza el principio de adición de vestigios de ruido adaptativo para reducir los errores principales del detector. La adaptación de ruido vestigial se divulga en la referencia [5], de Hong y otros.

Un inconveniente del EVRC VAD de banda repartida es que ocasionalmente toma malas decisiones y muestra una sensibilidad de frecuencia demasiado baja.

La detección de la actividad de voz la ha divulgado Freeman, véase la referencia [6], donde se divulga un VAD con espectro de ruido independiente, y Barret, véase la referencia [7], ha divulgado un mecanismo detector de tonos que no caracteriza equivocadamente el ruido de coches de baja frecuencia como tonos de señalización. Un inconveniente de las soluciones basadas en Freeman/Barret muestra ocasionalmente una sensibilidad demasiado baja (por ejemplo, para la música de fondo).

Otra detección de la actividad de la voz ha sido divulgada por Jenilek y otros, véase la referencia [10].

### Sumario

Un objeto de la invención es proporcionar un detector de voz y un detector de actividad de la voz que es más sensible a la actividad de voz sin experimentar los inconvenientes de los dispositivos de la técnica anterior.

Este objeto se consigue con un detector de voz y un detector de actividad de la voz que utilizan un detector de voz en el que se utiliza una señal de entrada, dividida en señales sub-banda que representan n sub-bandas de frecuencias diferentes, para calcular una relación señal-ruido (SNR) para cada sub-banda. Se calcula un valor de la SNR en el dominio de potencias para cada sub-banda, y se calcula al menos uno de los valores de la SNR de la potencia utilizando una función de ponderación no lineal. Se forma un valor único basándose en los valores SNR de la potencia y se compara el valor único con un umbral dado para generar una decisión de actividad de la voz en un puerto de salida del detector de voz. Al introducir una función de ponderación no lineal para una o más sub-bandas, la importancia de las sub-bandas que es probable que introduzcan ruido de la decisión en la métrica de la decisión real, se reduce selectivamente por medio de la función no lineal introducida tras el cálculo de la SNR.

Otro objeto de la invención es proporcionar un método que proporciona un detector de voz que es más sensible a la actividad de voz, sin experimentar los inconvenientes de los dispositivos de la técnica anterior.

Este objeto se consigue con un método para reducir selectivamente la importancia de las sub-bandas adaptativamente, para un detector de suma de SNR de voz de sub-banda, donde una señal de entrada al detector de voz se divide en n sub-bandas de frecuencias diferentes. La suma de SNR está basada en una ponderación no

lineal aplicada a las señales que representan al menos una sub-banda antes de efectuar la suma de SNR.

Una ventaja de la presente invención es que se mantiene la calidad de la voz, o incluso se mejora bajo ciertas condiciones en comparación con las soluciones de la técnica anterior.

Otra ventaja es que la invención reduce la velocidad media en condiciones de ruido no estacionario, tal como las condiciones de murmullos, en comparación con las soluciones de la técnica anterior.

**Breve descripción de los dibujos**

La figura 1 muestra una solución de la técnica anterior para un VAD.

La figura 2 muestra una descripción detallada de un detector de voz, utilizado en el VAD descrito en conexión con la figura 1.

La figura 3 muestra un primer modo de realización de un detector de voz de acuerdo con la presente invención.

La figura 4 muestra un gráfico que ilustra el rendimiento en actividad de voz para diferentes VAD.

La figura 5 muestra un primer modo de realización de un VAD, de acuerdo con la presente invención.

La figura 6 muestra un segundo modo de realización de un VAD, de acuerdo con la presente invención.

La figura 7 muestra un gráfico que ilustra resultados subjetivos obtenidos por un test de escucha experta de Mushra para diferentes VAD.

La figura 8 muestra un codificador de habla que incluye un VAD de acuerdo con la invención.

La figura 9 muestra un terminal que incluye un VAD de acuerdo con la invención.

**Descripción detallada**

La figura 1 muestra un detector de actividad de la voz VAD 10, similar al VAD divulgado en la referencia [1] denominado AMR VAD 1, y la figura 2 muestra una descripción detallada de un detector principal de voz utilizado.

El VAD 10 divide la señal entrante “señal de entrada” en tramas de muestras de datos. Estas tramas de muestras de datos se dividen en “n” sub-bandas de frecuencias diferentes por medio de un analizador de sub-bandas (SBA) 11 que calcula también el correspondiente nivel de entrada “level[n]” para cada sub-banda. Estos niveles se utilizan después para estimar el nivel de ruido de fondo “bckr_est[n]” en un estimador de nivel de ruido (NLE) 12, para cada sub-banda, mediante el filtrado en paso bajo de las estimaciones de niveles para tramas sin voz. Así, el NLE genera una condición estimada de ruido o condición de señal de fondo, por ejemplo, música, utilizada en una detector principal de voz (PVD).El PVD 13 utiliza la información de niveles “level[n]” y el nivel de ruido de fondo estimado “bckr_est[n]” para cada sub-banda “n” para formar una decisión “vad_prim” sobre si la trama de datos en curso contiene o no datos de voz. La decisión “vad_prim” se utiliza en el NLE 12 para determinar tramas sin voz.

La operación básica del PVD 13, que se describe con más detalle con relación a la figura 2, es supervisar cambios en las relaciones de señal-ruido (SNR) de la sub-banda y los cambios suficientemente grandes se considera que son de habla. Esto se obtiene calculando una relación señal-ruido *snr[n]* en cada sub-banda utilizando una función “Calc. SNR” en el bloque 20.

$$snr[n] = \frac{level[n]}{bckr\_est[n]} \qquad (1)$$

El valor SNR calculado se convierte en potencia tomando el cuadrado del valor de la SNR calculada para cada sub-banda, que se calcula en el bloque 21, y se forma un valor combinado de SNR para *snr_sum* basado en todas las sub-bandas. La base del valor SNR combinado es el valor medio de todas las SNR de potencia de las sub-bandas formado por el bloque 22 de suma de la figura 2.

$$snr\_sum = \frac{1}{k}\sum_{n=1}^{k}(snr[n])^2, \qquad (2)$$

donde k es el número de sub-bandas, por ejemplo 9 sub-bandas, como se ilustra en la figura 2.

La decisión de actividad de voz principal “vad_prim” del PVD 13 puede formarse entonces comparando el “snr_sum” calculado con un valor umbral “vad_thr” en el bloque 23. El valor umbral “vad_thr” se obtiene a partir de un circuito de adaptación del umbral (TAC) 24, como se ilustra en la figura 2. El valor umbral “vad_thr” se ajusta de acuerdo con

el nivel de ruido de fondo obtenido mediante la suma de todos los niveles de ruido de fondo de las sub-bandas desde el NLE 12, para aumentar la sensibilidad (disminuir el umbral), y evitar las tramas que faltan que contienen los datos de voz, si el nivel de ruido de fondo es alto.

Los niveles de entrada calculados en el SBA 11 se proporcionan también a un estimador estacionario (STE) 16 que proporciona información "stat_rat" al NLE 12, cuya información indica la estabilidad a largo plazo del ruido de fondo. En el VAD 10 se puede proporcionar también un módulo de ruido vestigial (NHM) 14, donde el NHM 14 se utiliza para ampliar el número de tramas que el PVD ha detectado que contienen habla. El resultado es una decisión de actividad de voz modificada "vad_flag" que se utiliza en el sistema del códec de habla, como se describe en conexión con la figura 8. La decisión "vad_flag" se proporciona al códec 15 de habla para indicar que la señal de entrada contiene habla, y el códec 15 de habla proporciona señales de "tono" y de "inflexión" al NLE 12. La decisión "vad_prim" puede ser también retroalimentada al NLE 12. Los bloques funcionales denominados SBA 11, NLE 12, NHM 14, códec 15 de habla y STE 16 son muy conocidos por una persona experta en la técnica y no se describe por tanto con más detalle.

Un inconveniente del PVD descrito de la técnica anterior es que puede indicar actividad de voz para el ruido de fondo no estacionario, tal como el ruido de fondo de murmullos. Un objetivo de la presente invención es modificar el PVD de la técnica anterior para reducir ese inconveniente.

La figura 3 muestra un primer modo de realización de un detector de voz principal no lineal NL PVD 30, que incluye los mismos bloques funcionales descritos en conexión con la figura 2 y un bloque funcional 31 para cada sub-banda "n". El bloque funcional 31 proporciona una ponderación no lineal del valor SNR calculado desde el bloque funcional 20, que es la modificación que reduce el problema de la técnica anterior. Para este modo de realización, la función no lineal se implementa para producir la snr_sum resultante de la suma de las SNR por medio de:

$$snr\_sum = \frac{1}{k}\sum_{n=1}^{k}\begin{cases} 0 & si\ snr[n] < sign\_thresh \\ (snr[n])^2 & en\ otro\ caso \end{cases}, \qquad (3)$$

donde "k" es el número de sub-bandas (por ejemplo, k=9), snr[n] es la relación señal-ruido para la sub-banda "n" y "sign_thresh" es el valor umbral significativo de la función no lineal.

La función no lineal es fijar en cero (0) el valor SNR de cada valor SNR calculado inferior al "sign_thresh" y mantenerlo inalterado para otros valores de SNR. El "sign_thresh" umbral significativo se fija preferiblemente en un valor mayor que uno (sign_thresh>1), y más preferiblemente en dos o mayor (sign_thresh≥2). El valor de SNR se eleva al cuadrado para convertirlo al dominio de potencias, como es obvio para una persona experta en la técnica. Un valor de SNR de uno o mayor dará como resultado un correspondiente valor de potencia de SNR de uno o mayor. Sin embargo, hay otras posibilidades con respecto a la implementación de la función no lineal del bloque funcional 31 cuando se calcula la snr_sum a partir de la suma de las SNR, tal como:

$$snr\_sum = \frac{1}{k}\sum_{n=1}^{k}\begin{cases} (sign\_floor)^2 & si\ sign\_floor < snr[n] < sign\_thresh \\ (snr[n])^2 & en\ otro\ caso \end{cases}, \qquad (4)$$

donde "k" es el número de sub-bandas (por ejemplo, k = 9), "sign_floor" es el valor predeterminado, snr[n] es la relación señal-ruido de la sub-banda "n" y "sign_thresh" es el valor umbral significativo de la función no lineal.

El "sign_thresh" umbral significativo se fija preferiblemente como se ha mencionado anteriormente, es decir, mayor que uno (sign_thresh>1), y más preferiblemente en dos o mayor (sign_thresh≥2). El valor predeterminado "sign_floor" es preferiblemente inferior a 1 (sign_floor<1) y más preferiblemente inferior o igual cero como cinco (sign_floor≤0,5).

La mejora en el rendimiento de la actividad de voz para el habla con ruidos de murmullos de fondo está ilustrada en la figura 4, que muestra el rendimiento de diferentes VAD. El gráfico presenta el valor medio de la decisión de actividad de voz "Valor medio (vad_DTX)" por el módulo de DTX vestigial, descrito con más detalles en la figura 8, para diferentes VAD en función de tres niveles de entrada en dBov y diferentes valores de SNR en dB. El término dBov significa "sobrecarga de dB". Un nivel dBov de 0 significa que el sistema está justamente en el umbral de sobrecarga. Una muestra digital de 16 bits tiene un máximo de +32767, que se corresponde con 0 dB. -26dB significa que el tamaño máximo de la muestra es de 26 dB por debajo del máximo. Los VAD ilustrados son:

VAD1: marcado con una cruz indicada con 41 para el nivel de entrada de -16dB, 44 para el nivel de entrada de -26dB y 47 para el nivel de entrada de - 36dB.

EVRC VAD: marcado con un cuadrado indicado con 42 para un nivel de entrada de -16dB, 45 para el nivel de

entrada de -26dBov y 48 para el nivel de entrada de -36 dBov.

VAD 5 (que es un VAD que comprende un detector principal de voz 30 de acuerdo con la invención): marcado con un triángulo indicado con 43 para el nivel de entrada de -16dBov, 46 para el nivel de entrada de -26dBov y 49 para el nivel de entrada de - 36dBov.

Debe indicarse que la actividad media "Valor medio (vad_DTX)" para el VAD 5 es significativamente inferior en comparación con el VAD 1 para todos los niveles de entrada con un valor de SNR por debajo de infinito, y el "Valor medio (vad_DTX)" para VAD 5 es inferior en comparación con el EVRC VAD para todos los niveles de entrada con un valor de SNR de 10 dB. Además, el VAD5 y el EVRC VAD muestran igualmente una buena actividad media y son compatibles para otros valores de SNR.

Debe mencionarse que el umbral significativo de las diferentes sub-bandas puede ser idéntico, o puede ser diferente, como se ilustra a continuación:

$$snr\_sum = \frac{1}{k}\sum_{n=1}^{k}\begin{cases}(sign\_floor[n])^2 & si\ \ sign\_floor[n] < snr[n] < sign\_thresh[n] \\ (snr[n])^2 & en\ otro\ caso\end{cases}, \quad (5)$$

donde "k" es el número de sub-bandas (por ejemplo, k = 9), "sign_floor[n]" es un valor predeterminado para cada sub-banda "n", "snr[n]" es la relación señal-ruido de la sub-banda "n", y "sign_thresh[n]" es el valor umbral significativo de la función no lineal en cada sub-banda "n".

El uso de diferentes umbrales significativos en diferentes sub-bandas conseguirá un rendimiento optimizado en frecuencia para ciertos tipos de ruidos de fondo. Esto significa que el umbral significativo podría fijarse en 1,5 para la función no lineal en el bloque $31_1$ a $31_5$, y en 2,0 en el bloque funcional $31_6$ - $31_9$ sin apartarse del concepto inventivo.

En la figura 5, se describe un primer modo de realización de un VAD 50 de acuerdo con la invención, que tiene los mismos bloques funcionales que el VAD de la técnica anterior descritos en conexión con la figura 1, excepto que se utiliza un detector principal de voz no lineal NL PVD 51, que tiene un bloque funcional no lineal como se describe en conexión con la figura 3, en lugar del PVD de la técnica anterior. Se puede conectar una unidad de control opcional CU 52 en el VAD 50, para hacer los ajuste del valor umbral significativo "sign_thresh" y del valor predeterminado "sign_floor" (si fuera posible) para cada sub-banda durante el funcionamiento. Los umbrales significativos son fijos, pero pueden cambiarse (actualizarse) por medio de la CU 52.

En la figura 5, el nivel de ruido de cada sub-banda se estima basándose en las señales de tono y de inflexión del códec 15 de habla, en las decisiones de vad_prim anteriores almacenadas en un registro de memoria accesible para el NLE 12 y en el valor estacionario del nivel stat_rat obtenido desde el STE 16. La configuración detallada de la adaptación del nivel de ruido de la sub-banda se describe en TS 26.094, referencia [1]. El funcionamiento del detector principal de voz no lineal NL PVD se ha descrito anteriormente.

Los primeros modos de realización muestran cómo puede utilizarse el detector principal de voz no lineal para mejorar la funcionalidad, de manera que se reducen las decisiones activas falsas. Sin embargo, para ciertas condiciones de ruido de fondo estables y estacionarias, tales como el ruido del coche y el ruido blanco, debe haber un equilibrio cuando se fijan los umbrales significativos. Para resolver este problema, el umbral significativo puede hacerse adaptativo basándose en un análisis independiente a plazo más largo de la condición del ruido de fondo.

Para condiciones en las que se supone una fuerte variación de energía de la sub-banda, se puede emplear un umbral significativo no estricto, y para condiciones en las que se supone una baja variación de la energía de la sub-banda se puede utilizar un umbral significativo más exigente. La adaptación del umbral significativo se diseña preferiblemente de manera que las partes activas de la voz no se usen en la estimación de la condición del ruido de fondo.

La figura 6 muestra un segundo modo de realización de un VAD 60 de acuerdo con la invención, provisto de un detector principal de voz no lineal NL PVD 61, cuyo valor umbral significativo de cada sub-banda en el bloque funcional no lineal, puede ser ajustado adaptativamente. Hay un detector de voz optimista OVD 62, con un ajuste de umbral significativo optimista fijo, que funciona continuamente en paralelo con el NL PVD 61 para producir una decisión optimista de la actividad de voz "vad_opt". El umbral significativo del NL PVD se adapta utilizando información del tipo de ruido de fondo que es analizada durante periodos de habla no activos indicados por "vad_opt" en un adaptador NCA 63 de la condición de ruido. Basándose en dos módulos adicionales, es decir, el OVD 62 y el NCA 63, el umbral significativo sign_thresh del NL PVD 61 se ajusta por medio de una señal de control del NCA 63. El detector de voz optimista OVD 62 es preferiblemente una copia del NL PVD 61 con un ajuste optimista (o agresivo) de un valor del umbral significativo, preferiblemente un valor fijo SF. Un valor preferido para el SF es 2,0.

La información del tipo de ruido de fondo, sobre la cual el NBA 63 genera la señal de control, es preferiblemente la

señal stat_rat generada en el STE 16, como se indica con la línea continua 64, pero la señal de control puede estar basada en otros parámetros que caracterizan el ruido, especialmente parámetros disponibles en el VAD 1 del TS 26.094 y a partir del análisis del códec de habla, como se indica con la línea de puntos 65, es decir, el valor de correlación de la tonalidad filtrada en paso alto, el señalizador de tono, o la variación del parámetro ptich_gain del códec de habla.

En el modo de realización preferido, el valor de stat_rat del STE 16 se utiliza como información tipo del ruido de fondo sobre el cual se basa la señal de control durante los periodos de habla no activos, como se indica con "vad_opt". Una modificación del algoritmo original descrito en TS 26.094 es que el cálculo del valor de estimación de la estacionalidad "stat_rat" se realiza continuamente en cada trama de decisión VAD. En el TS 26.094 de 3GPP, el cálculo de "stat_rat" se explica en la sección "3.3.5.2 Estimación del ruido de fondo".

La estacionalidad (stat_rat) se estima utilizando la ecuación siguiente:

$$stat\_rat = \sum_{n=1}^{9} \frac{MAX(STAT\_THR\_LEVEL, MAX(ave\_level_m[n], level_m[n]))}{MAX(STAT\_THR\_LEVEL, MIN(ave\_level_m[n], level_m[n]))}$$

donde $level_m$ es el vector de los niveles actuales de la amplitud de sub-banda y $ave\_level_m$ es una estimación del valor medio de niveles anteriores de sub-banda. STAT_THR_LEVEL se fija en un valor apropiado, por ejemplo en 184 (Escalamiento/precisión del VAD 1 de TS 26.094).

Un valor alto de "stat_rat" indica la existencia de grandes variaciones del nivel dentro de la banda, una valor bajo de "stat_rat" indica variaciones menores del nivel dentro de la banda.

La historia de las decisiones de vad_opt se almacena en un registro de memoria que es accesible para la NCA durante su funcionamiento.

La NCA añadida 63 utiliza el valor de "stat_rat" para ajustar el NL PVD 61 como sigue:

Cuando el vad_opt ha indicado inactividad de habla durante al menos 80 ms,

si el valor de "stat_rat" es más alto que un umbral STAT_THR (que indica alta variabilidad), generar una señal de control que desplace el "sign_thresh" de la ecuación (3) - (5) hacia el valor 2,0 con un tamaño del paso de 0,02,

si el valor "stat_rat" es inferior al umbral STAT_THR (que indica baja variabilidad), generar una señal de control que desplace el "sign_thresh" de la ecuación (3) - (5) hacia el valor 0,125 con un tamaño del paso de 0,01.

Si vad_opt indicase cualquier actividad de voz dentro de los últimos 80 ms, no generar ninguna señal de control para adaptar el valor de "sign_thresh" en la ecuación (3) - (5).

El resultado de la solución adaptativa descrita anteriormente es que el umbral (o umbrales) significativos son ajustados continuamente durante los supuestos periodos de inactividad, y el detector principal de voz NL-PVD se hace más (o menos) sensible al modificar el umbral (o umbrales) significativos dependiendo del análisis de energía de la sub-banda.

La figura 7 muestra resultados subjetivos obtenidos a partir de los tests de escucha experta de Mushra de material crítico, consistente en habla de -26 dBov en combinación con diferentes ruidos de fondo, tales como el coche, el garaje, murmullos, centros comerciales y calle (todos con una SNR de 10 dB). Para el test Mushra, las muestras de habla de diferentes codificadores se ordenan respecto a la calidad. El test utilizaba un modo AMR MR 122 como calidad de referencia alta, indicada como "Ref". Las funciones comparadas del VAD fueron codificadas utilizando el modo AMR MR59 y consistía en un VAD 1, EVRC VAD (utilizado sin supresión de ruido) y el VAD divulgado con umbrales significativos fijos de 2,0 y un suelo significativo de 0,5, indicado como VAD5.

En la figura 7 se indican los intervalos de un 95% de confianza para VAD diferentes y, desde el punto de vista de la escucha, no hay diferencia esencial entre los diferentes VAD, aunque la actividad media para la presente invención (VAD5) es considerablemente inferior en comparación con el VAD1, véase la figura 4.

La figura 8 muestra un sistema completo 80 de codificación que incluye un detector de actividad de la voz VAD 81, diseñado preferiblemente de acuerdo con la invención, y un codificador 82 de habla que incluye Transmisión Discontinua/Ruido de Confort (DTX/CN). La figura 8 muestra un codificador 82 de habla simplificado, cuya descripción detallada puede encontrarse en las referencias [8] y [9]. El VAD 81 recibe una señal de entrada y genera un "vad_flag" de decisión. El codificador 82 de habla comprende un módulo 83 de DTX vestigial que puede añadir siete tramas extra al "vad_flag" recibido desde el VAD 81; para más detalles ver la referencia [9]. Si "vad_DTX" = "1", se detecta voz, y si "vad_DTX" = "0", no se detecta voz. La decisión de "vad_DTX" controla un interruptor 84 que está fijado en la posición 0 si "vad_DTX" es "0" y en posición 1 si "vad_DTX" es "1".

"vad_DTX" es reenviado también en este ejemplo a un códec 85 de habla conectado a la posición 1 del interruptor

84, el códec 85 de habla usa el “vad_DTX” junto con la señal de entrada para generar el “tono” y la “inflexión” al VAD 81, como se ha descrito anteriormente. También es posible reenviar el “vad_flag” desde el VAD 81 en lugar del “vad_DTX”. El “vad_flag” es reenviado a una memoria intermedia de ruido de confort (CNB) 86 que sigue el rastro de las últimas siete tramas de la señal de entrada. Esta información es reenviada a un codificador 87 de ruido de confort (CNC) que recibe también el “vad_DTX” para generar ruido de confort durante las tramas sin voz; para más detalles ver la referencia [8]. El CNC se conecta a la posición 0 del interruptor 84.

La figura 9 muestra un terminal 90 de usuario, de acuerdo con la invención. El terminal comprende un micrófono 91 conectado a un dispositivo 92 de A/D para convertir la señal analógica en señal digital. La señal digital es alimentada a un codificador 93 de habla y al VAD 94, como se describe en conexión con la figura 8. La señal del codificador de habla es reenviada a una antena ANT, a través de un transmisor TX y un filtro dúplex DPLX, y es transmitida desde ahí. La señal recibida en la antena ANT es reenviada a una rama de recepción RX, a través del filtro dúplex DPLX. Las operaciones conocidas de la rama de recepción RX son llevadas a cabo para el habla recibida en la recepción, y se repiten a través del altavoz 95.

La señal de entrada al detector de voz descrito anteriormente ha sido dividida en sub-señales, donde cada una de ellas representa una sub-banda de frecuencias. La sub-señal puede ser un nivel de entrada calculado para una sub-banda, pero también es concebible crear una sub-señal basada en el nivel de entrada calculado, por ejemplo, convirtiendo el nivel de entrada al dominio de potencias, multiplicando el nivel de entrada por sí mismo antes de ser alimentada al detector de voz. Las sub-señales que representan las sub-bandas de frecuencias pueden generarse también mediante auto-correlación, como se describe en las referencias [2] y [4], donde las sub-señales se expresan en el dominio de potencias sin que sea necesaria ninguna conversión. Lo mismo es aplicable a las sub-señales de fondo recibidas en el detector de voz.

Declaraciones relativas a la invención:

- El detector de voz en cuanto a ruido estimado o condición de señal de fondo, está basado en partes no activas de voz de la señal de entrada.

- El detector de voz en el sentido de detector de voz, está configurado para sustituir cada valor de SNR (snr[n]) inferior al valor del umbral significativo específico de la sub-banda (sign_thresh) por un valor predeterminado en la función no lineal. Donde dicho valor predeterminado es cero (0) o el valor predeterminado es inferior al valor SNR de cada sub-banda.

El valor predeterminado podría ser especificado también como menor que uno (sign_floor < 1), preferiblemente menor o igual a cero coma cinco (sign_floor $\leq$ 0,5).

- El detector de actividad de la voz, en el sentido de detector principal de voz (30; 51; 61) está provisto de una memoria en la cual son almacenadas las decisiones previas de la actividad de voz (vas_prim); y el ruido de fondo estimado calculado en el estimador (12) de nivel de ruido de cada sub-banda, está basado además en la decisión previa almacenada de la actividad de voz principal (vad_prim).

- El detector de actividad de la voz comprende además:

- medios (62, 63) para producir una señal de control basada en parámetros que caracterizan el ruido en la señal de entrada, utilizándose dicha señal de control en el detector principal de voz (61) para ajustar selectivamente un umbral significativo específico de la sub-banda (sign_thresh) en la función no lineal.

- Comprendiendo además el detector de actividad de la voz un estimador estacionario (16) configurado para producir un valor de estacionalidad (stat_rat) basado en el nivel de entrada calculado (level[n]) para cada sub-banda, donde dicha señal de control está basada en el valor de estacionalidad (stat_rat).

- El detector de actividad de la voz, en el que dichos medios para producir una señal de control comprende un detector de voz secundario (62), como se define en cualquiera de las reivindicaciones 1 - 20, configurado para producir una decisión de la actividad de voz secundaria (vad_opt), estando basada además dicha señal de control (sign_thresh) en la decisión de la actividad de voz secundaria (vad_opt).

- El detector de actividad de la voz, en el que el detector de voz secundario (62) usa una función no lineal que tiene un umbral significativo fijo (SF) para todas las sub-bandas.

**Abreviaturas**

AMR Velocidad múltiple adaptativa

ANT Antena

CNB Memoria intermedia del ruido de confort

CNC Codificador del ruido de confort

| | |
|---|---|
| DTX | Transmisión discontinua |
| DPLX | Filtro dúplex |
| EVRC | Velocidad variable reforzada (IS - 127) |
| NCA | Adaptador de condición de ruido |
| NHM | Módulo de ruido vestigial |
| NLE | Estimador de nivel de ruido |
| NL PVD | Detector de voz principal no lineal |
| OVD | Detector de voz optimista |
| PVD | Detector de voz principal |
| RX | Rama de recepción |
| SBA | Analizador de sub-banda |
| SNR | Relación señal-ruido |
| STE | Estimador de estacionalidad |
| TAC | Circuito de adaptación de umbral |
| TX | Transmisor |
| VAD | Detector de actividad de voz |

**Referencias**

[1] “Adaptive Multi Rate (AMR) speech codec (Códec de habla de velocidad múltiple adaptativa; Voice Activity Detector (VAD) (Detector de Actividad de Voz” 3GPP TS 26.094 V6.0.0 (2004-12)

[2] “Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems” (Códec de velocidad variable reforzado, Opción 3 de servicio de habla para sistemas digitales de espectro extendido de banda ancha), 3GPP2.C.S0014-A v 1.0, 2004-05

[3] US 5.963.901 A1, de Vähätalo, con el título “Method and Device for voice activity detection, and a communication Device” (Método y dispositivo para la detección de actividad de la voz y dispositivo de comunicaciones), asignado a Nokia, 10 de Diciembre de 1996.

[4] US 5.742.734 A1, de De Jaco, con el título “Encoding rate selection in a variable rate vocoder” (Selección de la velocidad de codificación en un codificador de voz de velocidad variable), asignado a Qualcomm, 10 de Agosto de 1994.

[5] US 5.410.632 A1, de Hong, con el título “Variable hangover time in a voice activity detector” (Variabilidad vestigial de tiempo en un detector de actividad de voz), asignado a Motorola, 23 de Diciembre de 1991.

[6] US 5.276.765 A1, de Freeman, con el título “Voice activity detection” (Detección de actividad de voz), 10 de Marzo de 1989.

[7] US 5.749.067 A1, de Berret, con el título “Voice activity detector” (Detector de actividad de voz), 8 de Marzo de 1996.

[8] “Adaptive Multi-rate (AMR) speech codec; Comfort Noise AMR Speech Traffic Channels” (Códec de habla adaptativo de múltiples velocidades (AMR); Canales de tráfico de habla de ruido de confort AMR), 3GPP TS 26.094, V6.0.0 (2004-12).

[9] Adaptive Multi-rate (AMR) speech codec; Source Control Rate Operation” (Códec de habla adaptativo de múltiples velocidades (AMR); Funcionamiento de la velocidad de control de la fuente), 3GPP TS 26.093, V6.1.0 (2006-06).

[10] Jelinek M et al, Advances in source-controlled variable bit rate wideband speech coding. Special WS en MAW (SWIM); (Jelinek y otros, Avances en codificación del habla de banda ancha con velocidad de bits variable controlada por la fuente. WS Especial en MAW (SWIM). Conferencias de expertos en proceso del habla, Enero de 2004, páginas 1 - 8.

**REIVINDICACIONES**

1. Un detector de voz (30; 51; 61) que responde a una señal de entrada que se divide en sub-señales, representando cada una de ellas una sub-banda (n) de frecuencias, donde dicho detector de voz comprende:

- un primer puerto de entrada configurado para recibir dichas sub-señales,

- un segundo puerto de entrada configurado para recibir una sub-señal de fondo basada en dichas sub-señales y

- medios para calcular (20), para cada sub-banda, un valor SNR (snr[n]) basado en la correspondiente sub-señal y en la sub-señal de fondo;

**caracterizado porque** dicho detector de voz (30; 51; 61) comprende además:

- medios para calcular (31n, 21) un valor de SNR de potencia para cada sub-banda,

donde al menos uno de dichos valores de SNR de potencia se calcula basándose en una función de ponderación no lineal

- medios para formar (22) un valor único (snr_sum) basado en los valores de potencia calculados, y

- medios para comparar (23) dicho valor único (snr_sum) con un valor umbral dado (vad_thr) para tomar una decisión de actividad de voz (vad_prim) presentado en un puerto de salida.

2. El detector de voz según la reivindicación 1, en el que cada uno de dichos valores de SNR de potencia se calcula basándose en una función de ponderación no lineal.

3. El detector de voz según la reivindicación 1 o la reivindicación 2, en el que el detector de voz está configurado para aplicar la función de ponderación no lineal al valor SNR, antes de calcular el valor de la SNR de la potencia.

4. El detector de voz según cualquiera de las reivindicaciones 1 - 3, en el que el detector de voz está configurado para usar un valor umbral significativo específico de la sub-banda (sign_thresh) en la función de ponderación no lineal, para suprimir selectivamente las sub-bandas.

5. El detector de voz según la reivindicación 4, en el que el valor umbral significativo específico de la sub-banda (sign_thresh) es diferente para al menos dos sub-bandas.

6. El detector de voz según la reivindicación 4, en el que el valor umbral significativo específico de la sub-banda (sign_thresh) es el mismo para todas las sub-bandas.

7. El detector de voz según cualquiera de las reivindicaciones 4 - 6, en el que el valor umbral significativo específico de la sub-banda tiene un valor mayor que uno (sign_thresh > 1), preferiblemente dos o mayor (sign_thresh $\geq$ 2).

8. El detector de voz según cualquiera de las reivindicaciones 4 - 7, en el que el detector de voz está configurado para tener un valor umbral significativo fijo específico de la sub-banda.

9. El detector de voz según cualquiera de las reivindicaciones 4 - 7, en el que el detector de voz está configurado para ajustar adaptativamente el valor umbral significativo específico de la sub-banda, basándose en el ruido estimado o en la condición de la señal de fondo.

10. El detector de voz según cualquiera de las reivindicaciones 4 - 9, en el que el detector de voz está configurado para sustituir cada valor SNR (snr[n]) que sea menor que el valor umbral significativo fijo específico de la sub-banda (sign_thresh) por un valor predeterminado en la función de ponderación no lineal.

11. El detector de voz según cualquiera de las reivindicaciones 1 - 10, en el que dicha sub-señal de fondo para cada sub-banda se calcula basándose en decisiones anteriores de la actividad de voz principal (vad_prim) calculados en el detector de voz (51, 61).

12. El detector de voz según cualquiera de las reivindicaciones 1 - 11, en el que la señal de entrada contiene nueve sub-bandas de frecuencias.

13. El detector de voz según cualquiera de las reivindicaciones 1 - 12, en el que los medios para calcular los valores SNR de potencia para cada sub-banda están basados además en una función cuadrática implementada en un convertidor (21).

14. El detector de voz según cualquiera de las reivindicaciones 1 - 13, en el que los medios para formar un valor único (snr_sum) comprenden un bloque (22) de suma en el cual se forma el valor medio de todas las SNR de potencia de las sub-bandas.

15. El detector de voz según cualquiera de las reivindicaciones 1 - 14, en el que el detector de voz comprende además un circuito (24) adaptador de umbral, que produce dicho valor umbral (vad_thr) como respuesta a una señal (nivel de ruido) generada mediante la suma de la sub-señal de fondo para todas las sub-bandas.

16. El detector de voz según cualquiera de las reivindicaciones 1 - 15, en el que cada sub-señal está basada en un nivel de entrada calculado (level[n]) para cada sub-banda, y cada sub-señal de fondo está basada en un nivel de ruido de fondo estimado (bckr_est[n]) para cada sub-banda.

17. Un detector de actividad de la voz (50; 60; 81; 94) utilizado para determinar si hay datos de voz contenidos en una señal de entrada, **caracterizado porque** dicho detector de actividad de la voz (50; 60; 81; 94) comprende un detector de voz principal (30; 51; 61) como se define en cualquiera de las reivindicaciones 1 - 16.

18. El detector de actividad de la voz de acuerdo con la reivindicación 17, que comprende además:

- un analizador (11) de sub-bandas configurado para dividir dicha señal de entrada en tramas de muestras de datos, y para dividir además las tramas de muestras de datos en sub-bandas de frecuencias, configurado además dicho analizador de sub-bandas para calcular un correspondiente nivel de entrada (level[n]) para cada sub-banda, y

- un estimador (16) de nivel de ruido configurado para generar una estimación del nivel de ruido de fondo (bckr_est[n]) para cada sub-banda, basándose en los niveles de entrada (level[n]) calculados.

19. Un nodo de un sistema de telecomunicaciones que comprende un detector de actividad de la voz como se define en cualquiera de las reivindicaciones 17 - 18.

20. El nodo según la reivindicación 19, en el que el nodo es un terminal (90).

21. Un método de detección de voz de sub-banda de suma de SNR para suprimir selectivamente sub-bandas del detector de voz de sub-banda de suma de SNR, **caracterizado porque** dicha suma de SNR está basada en una ponderación no lineal para al menos una sub-banda, antes de sumar las SNR.

22. El método según la reivindicación 21, en el que se efectúa una ponderación no-lineal para cada una de dichas sub-bandas, antes de sumar las SNR.

23. El método según cualquiera de las reivindicaciones 21 - 22, en el que el método comprende calcular un valor de SNR de potencia para cada sub-banda, antes de sumar las SNR.

24. El método según cualquiera de las reivindicaciones 21 - 23, en el que la ponderación no lineal está basada en una función no lineal:

$$snr\_sum = \frac{1}{k}\sum_{n=1}^{k}\begin{cases}(sign\_floor[n])^2 & si\ \ sign\_floor[n] < snr[n] < sign\_thresh[n] \\ (snr[n])^2 & en\ otro\ caso\end{cases}$$

snr_sum es el resultado de la suma de las SNR,

k es el número de sub-bandas de frecuencias,

sign_floor es un valor predeterminado,

snr[n] es la relación señal-ruido de la sub-banda “n”, y

sign_thresh es el valor umbral significativo de la función de ponderación no lineal.

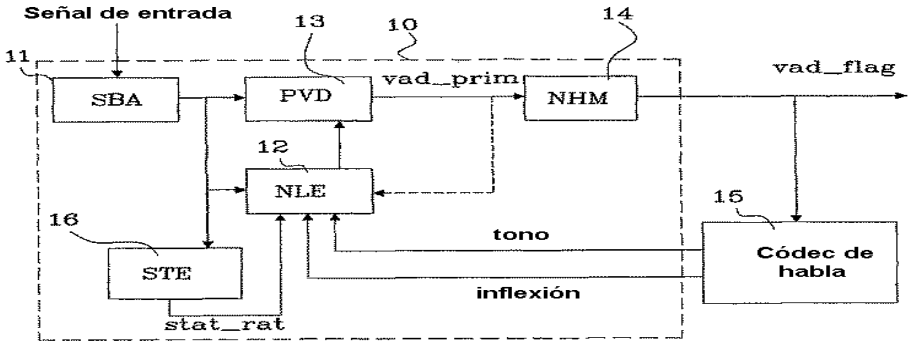


Fig. 1 (Técnica anterior)

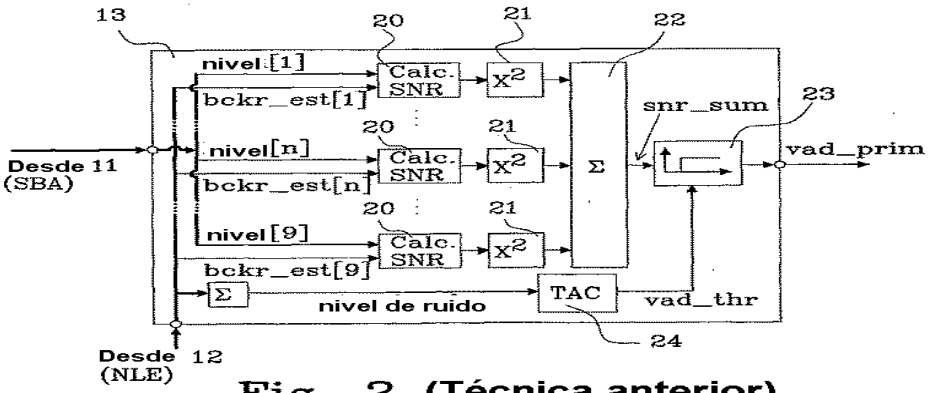


Fig. 2 (Técnica anterior)

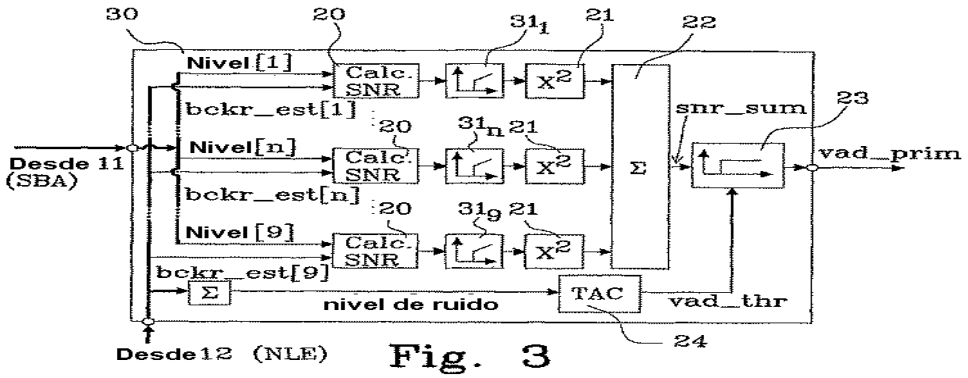
30
20
31 1
21
22
Nivel [1]
Calc. SNR
X2
bckr_est[1]
Nivel[n]
31 n
Σ
snr_sum
23
vad_prim
Desde 11 (SBA)
bckr_est[n]
Nivel [9]
31 9
bckr_est[9]
nivel de ruido
TAC
vad_thr
24
Desde 12 (NLE)

Fig. 3

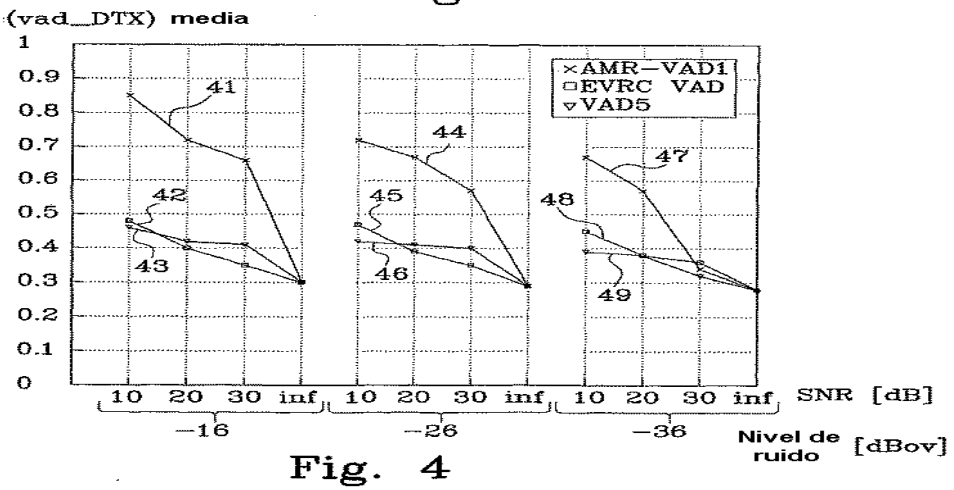
(vad_DTX) media
AMR-VAD1
EVRC VAD
VAD5
41
42
43
44
45
46
47
48
49
SNR [dB]
Nivel de ruido [dBov]

Fig. 4

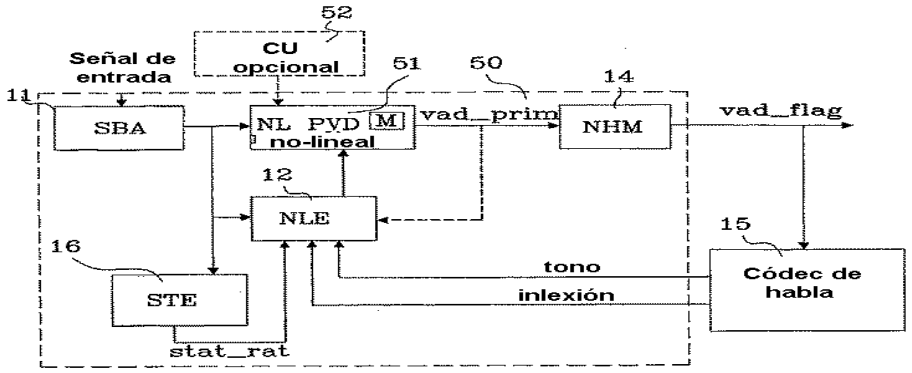
52
CU
opcional
Señal de
entrada
51
50
14
11
SBA
NL PVD M
no-lineal
vad_prim
NHM
vad_flag
12
NLE
15
16
tono
Códec de
habla
STE
inlexión
stat_rat

Fig. 5

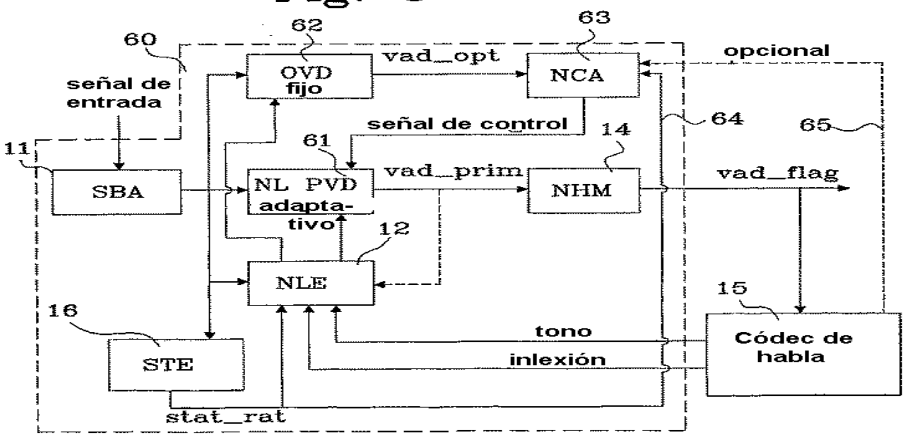
60
62
63
opcional
vad_opt
OVD
fijo
NCA
señal de
entrada
señal de control
64
65
11
61
14
SBA
NL PVD
adapta-
tivo
vad_prim
NHM
vad_flag
12
NLE
15
16
tono
Códec de
habla
STE
inlexión
stat_rat

Fig. 6

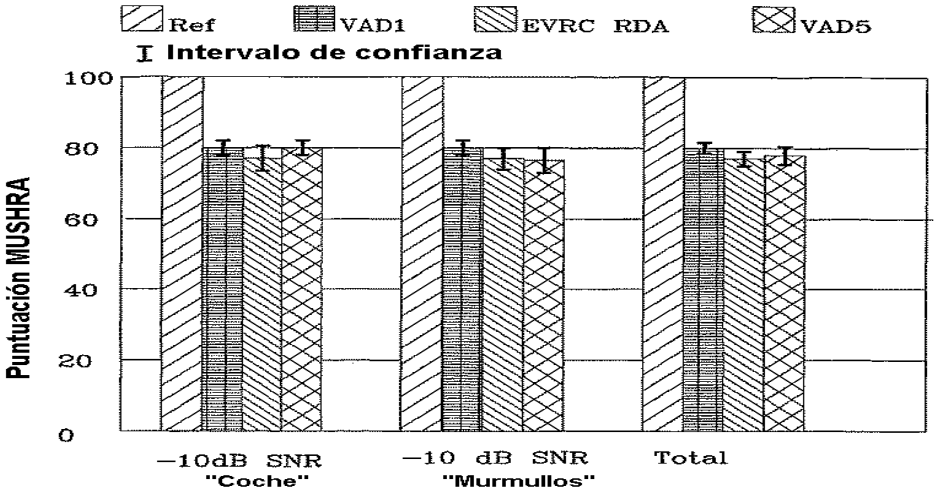
Ref
VAD1
EVRC RDA
VAD5
I Intervalo de confianza
Puntuación MUSHRA
100
80
60
40
20
0
−10dB SNR
"Coche"
−10 dB SNR
"Murmullos"
Total

Fig. 7

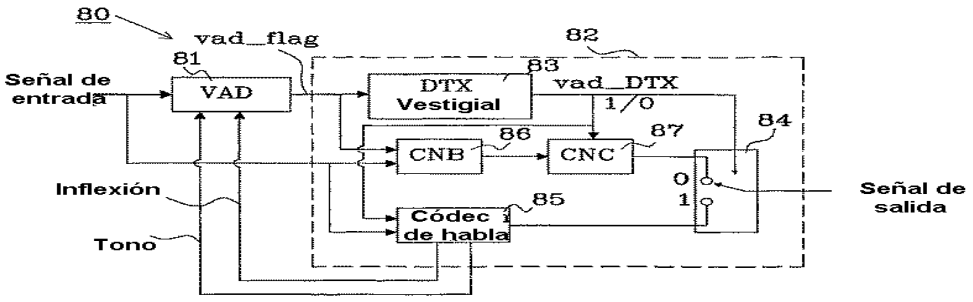
80
vad_flag
82
81
83
Señal de entrada
VAD
DTX Vestigial
vad_DTX
1/0
84
86
87
CNB
CNC
0
1
Señal de salida
Inflexión
85
Codec de habla
Tono

Fig. 8

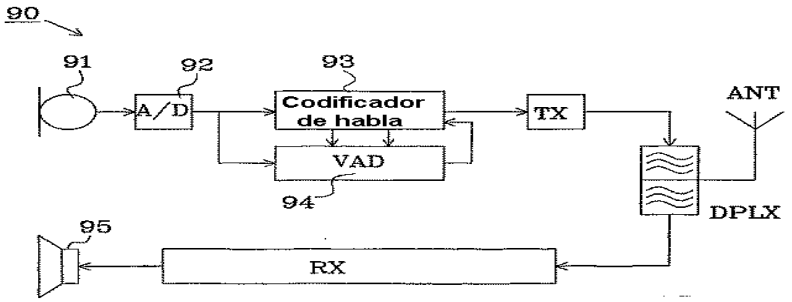
90
91
92
93
A/D
Codificador de habla
TX
ANT
VAD
94
DPLX
95
RX

Fig. 9