(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 430 121**

(51) Int. CI.:

G10L 21/0208 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **01.06.2012 E 12170407 (6)**

(97) Fecha y número de publicación de la concesión europea: **10.07.2013 EP 2530673**

(54) Título: **Equipo de audio que comprende unos medios de supresión de ruido de una señal de habla mediante filtrado de retardo fraccionario**

(30) Prioridad:

01.06.2011 FR 1154825

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
19.11.2013

(73) Titular/es:

PARROT (100.0%)
174 quai de Jemmapes
75010 Paris, FR

(72) Inventor/es:

VITTE, GUILLAUME y
HERVE, MICHAEL

(74) Agente/Representante:

FÀBREGA SABATÉ, Xavier

DESCRIPCIÓN

Equipo de audio que comprende unos medios de supresión de ruido de una señal de habla mediante filtrado de retardo fraccionario

La invención se refiere al tratamiento del habla en un medio ruidoso.

Se refiere, en particular, al tratamiento de las señales de habla captadas por unos dispositivos de telefonía de tipo “manos libres” destinados a ser utilizados en un entorno ruidoso.

Estos aparatos incorporan uno o varios micrófonos (“micros”) sensibles, que captan no solo la voz del usuario sino igualmente el ruido circundante, ruido que constituye un elemento perturbador que puede llegar en ciertos casos a convertir en ininteligible el habla del locutor. Lo mismo sucede si se quieren poner en práctica técnicas de reconocimiento de voz, pues es difícil efectuar un reconocimiento de forma sobre hablas ahogadas en un nivel de ruido elevado.

Esta dificultad unida a los ruidos circundantes es particularmente apremiante en el caso de los dispositivos “manos libres” para vehículos automóviles, ya se trata de equipos incorporados al vehículo o bien de accesorios en forma de carcasa inmóvil que integra todos los componentes y funciones de tratamiento de la señal para la comunicación telefónica.

En efecto, la importante distancia entre el micro (colocado al nivel del salpicadero o en un ángulo superior del techo del habitáculo) y el locutor (cuyo alejamiento está condicionado por la posición de la conducción) provoca la captación de un nivel de ruido relativamente elevado, que hace difícil la extracción de la señal útil, ahogada en el ruido. Así mismo, el medio muy ruidoso típico del entorno del automóvil presenta unas características espectrales no fijas, es decir que evolucionan de manera imprevisible en función de las condiciones de la conducción: paso por calzadas bacheadas o adoquinadas, la radio del vehículo en funcionamiento, etc.

Dificultades del mismo tipo se presentan en el caso de que el dispositivo consista en unos cascos de audio de tipo micro/cascos combinado utilizado para funciones de comunicación como por ejemplo funciones de telefonía “manos libres”, como complemento de la escucha de una fuente de audio (música, por ejemplo) proveniente de un aparato al que están conectados los cascos.

En este caso, se trata de utilizar una inteligibilidad suficiente de la señal captada por el micro, es decir de la señal de habla del locutor próximo (el portador de los cascos), o bien los cascos pueden ser utilizados en un entorno ruidoso (metro, calle de mucho tránsito, tren, etc.), de manera que el micro captará no solo el habla del portador de los cascos, sino los ruidos parásitos circundantes. El portador está ciertamente protegido de este ruido por los cascos, en especial si se trata de un modelo con auriculares cerrados que aíslen el ruido del exterior, y todavía más si los cascos están provistos de un “control activo del ruido”. Por contra, el locutor distante (el que se encuentra en el otro extremo del canal de comunicación) sufrirá ruidos parásitos captados por el micro y que vienen a interponerse y a interferir con la señal de habla del locutor próximo (el portador de los cascos). En particular, determinados formantes del habla esenciales para la comprensión de la voz quedan a menudo ahogados en componentes de ruido que habitualmente se encuentran en los entornos habituales.

La invención se refiere, más en concreto, a técnicas de supresión de ruido que incorporan varios micros, generalmente dos micros, para combinar de manera equilibrada las señales captadas simultáneamente por estos micros con el fin de aislar los componentes del habla útiles de los componentes de ruidos parásitos.

Una técnica clásica consiste en colocar y orientar uno de los micros para que capte principalmente la voz del locutor, mientras que el otro se dispone para que capte un componente de ruido más importante que el micro principal. La comparación de los signos captados permite extraer la voz del ruido ambiental mediante el análisis de la coherencia espacial de las dos señales, con medios software relativamente simples.

El documento US 2008/0280653 A1 describe una configuración de este tipo, en la que uno de los micros (el que capta principalmente la voz) es el de un auricular inalámbrico que lleva el conductor del vehículo, mientras que el otro (el que capta principalmente el ruido) es el del aparato telefónico, situado a distancia dentro del habitáculo del vehículo, por ejemplo acoplado al salpicadero.

Esta técnica, sin embargo, tiene el inconveniente de que se necesitan dos micros distantes, de forma que la eficacia es tanto más elevada cuanto más alejados están los dos micros. Debido a ello, esta técnica no es aplicable al dispositivo en el que los dos micros están próximos, por ejemplo dos micros incorporados en el frontal de una radio de vehículo automóvil, o dos micros que estuvieran dispuestos sobre una de las carcasas de un auricular de los cascos de audio.

Otra técnica más, llamada conformación de haces, consiste en crear mediante medios software una directividad que mejore la relación señal/ruido de la red o “antena” de micros. El documento US 2007/0165879 A1 describe una técnica de este tipo, aplicada a un par de micros no direccionales colocados de espaldas. Un filtrado adaptativo de las señales captadas permite derivar de salida una señal en la que el componente de voz ha sido reforzado.

No obstante, se considera que un método de este tipo no proporciona buenos resultados más que a condición de que disponga de al menos ocho micros, resultando en prestaciones extremadamente limitadas cuando solamente se utilizan dos micros.

El problema general de la invención es, en un contexto como el referido, proceder a una reducción eficaz del ruido que permita transmitir al locutor distante una señal vocal representativa del habla emitida por el locutor próximo (conductor del vehículo o portador de los cascos), liberando a esta señal de los componentes parásitos del ruido exterior existentes en el entorno de este locutor próximo.

El problema de la invención es igualmente, en tal situación, el de poder incorporar a la vez un conjunto de micros de un número reducido (de modo ventajoso, dos micros solamente) y relativamente próximos (típicamente una separación de solo algunos centímetros). Otro aspecto importante del problema es la necesidad de restituir una señal de habla natural e inteligible, es decir no distorsionada y cuyo espectro de frecuencias útiles no resulte cercenado por los tratamientos de supresión de ruido.

Con este fin la invención propone un equipo de audio del tipo general divulgado por el documento US 2008/0280653 A1 precitado, es decir que comprende: un conjunto de dos sensores microfónicos aptos para recoger el habla del usuario del equipo y para emitir unas señales de habla ruidosas respectivas; unos medios de muestreo de las señales de habla emitidas por los sensores microfónicos; y unos medios de supresión de ruido de una señal de habla, que reciben como salida las muestras de las señales de habla emitidas por los dos sensores microfónicos, y emiten de salida una señal de habla sin ruidos representativa del habla emitida por el usuario del equipo. Los medios de supresión de ruido son unos medios de reducción de ruido no frecuencial que comprenden un combinador con filtro adaptativo de las señales emitidas por los dos sensores microfónicos, que operan mediante la búsqueda iterativa que tiene por objeto anular el ruido captado por uno de los sensores microfónicos en base a una referencia de ruido dada por la señal emitida por el otro sensor microfónico.

Como característica distintiva de la invención, el filtro adaptativo es un filtro de retardo fraccionario, apto para modelar un retardo inferior al periodo de muestreo de los medios de muestreo. El equipo comprende además unos medios de detección de la actividad vocal aptos para emitir una señal representativa de la presencia o ausencia de habla por el usuario del equipo, y el filtro adaptativo recibe igualmente como entrada la señal de presencia o ausencia de habla, para, de forma selectiva: i) o bien operar una búsqueda adaptativa de los parámetros de filtro en ausencia de habla, ii) o bien congelar estos parámetros del filtro en presencia de habla.

El filtro adaptativo es, en especial, apto para estimar un filtro óptimo H, como:

$$\hat{H} = \hat{G} \otimes \hat{F}$$

con:

$$x'(n) = G \otimes x(n) \qquad \textbf{y} \qquad G(k) = \mathrm{sinc}(k + \tau / Te)\,,$$

representando	$\hat{H}$	la estimación del filtro óptimo H, la transferencia de ruido entre los dos sensores microfónicos para una respuesta de impulso incluyendo un retardo fraccionario,
representando	$\hat{G}$	la estimación del filtro del retardo fraccionario G entre los dos sensores microfónicos,
representando	$\hat{F}$	la estimación de la respuesta acústica del entorno,
indicando	$\otimes$	una convolución,
siendo	$x(n)$	la serie de muestras de la señal de entrada del filtro H,
siendo	$x'(n)$	la serie x(n) desplazada el retardo τ,
siendo	Te	el periodo de muestreo de la señal de entrada del filtro H,
siendo	τ	dicho retardo fraccionario, igual a un submúltiplo de Te, e
indicando	sinc	la función seno cardinal.

De modo preferente, el filtro adaptativo es un filtro para algoritmo de predicción lineal de tipo mínimos cuadrados medios, LMS.

En una forma de realización, el equipo comprende una cámara de vídeo dirigida hacia el usuario del equipo y apta para captar una imagen de éste, y los medios de detección de actividad vocal comprenden unos medios de análisis de vídeo aptos para analizar la imagen producida por la cámara y para emitir, como respuesta, dicha señal de presencia o de ausencia de habla por dicho usuario.

En otra forma de realización, el equipo comprende un sensor fisiológico apto para situarse en contacto con la cabeza

del usuario del equipo para quedar acoplado a ella con el fin de captar las vibraciones vocales no acústicas transmitidas por conducción ósea interna, y los medios de detección de actividad vocal comprenden unos medios aptos para analizar la señal emitida por el sensor fisiológico y para emitir, como respuesta, dicha señal de presencia o de ausencia de habla por dicho usuario, en especial mediante la evaluación de la energía de la señal emitida por el sensor fisiológico y su comparación con un umbral.

El equipo puede en particular ser unos cascos de audio del tipo combinado micro/cascos, que comprenda: unos auriculares cada uno de los cuales incorpore un transductor de reproducción sonora de una señal de audio alojada en una carcasa provista de una almohadilla circumaural; dichos dos sensores microfónicos, dispuestos sobre la carcasa de uno de los auriculares; y dicho sensor fisiológico incorporado a la almohadilla de uno de los auriculares y situado en una región de éste apta para situarse en contacto con la mejilla o con la sien del portador de los cascos. Estos dos sensores microfónicos están, de modo preferente, alineados en una red lineal siguiendo una dirección principal dirigida hacia la boca del usuario del equipo.

A continuación se describirá un ejemplo de puesta en práctica del dispositivo de la invención, con referencia a los dibujos adjuntos, en los que las mismas referencias numéricas designan a lo largo de ellos elementos idénticos o funcionalmente similares.

La Figura 1 ilustra de manera esquemática, en forma de bloques funcionales, la manera en la que se lleva a cabo el tratamiento de la supresión de ruido según la invención.

La Figura 2 es una representación gráfica de la función seno cardinal modelada en el tratamiento de la supresión de ruido de la invención.

Las Figuras 3a y 3b son dos representaciones de la función seno cardinal de la Figura 2, respectivamente para los diferentes puntos de una serie de muestras de señal, y para la misma serie desplazada en el tiempo un valor fraccionario.

La Figura 4 es una representación de la respuesta acústica del entorno con, como ordenada, la amplitud y, como abscisa, los coeficientes del filtro que representan esta transferencia.

La Figura 5 es análoga a la Figura 4, después de la convolución con una respuesta de seno cardinal.

La Figura 6 es una representación esquemática de una forma de realización consistente en la utilización de una cámara para asegurar la detección de actividad vocal.

La Figura 7 ilustra de forma general un conjunto de micros/cascos combinado al cual pueden aplicarse las enseñanzas de la invención.

La Figura 8 es un esquema de conjunto que ilustra en forma de bloques funcionales la manera en la que puede llevarse a cabo el tratamiento de la señal para emitir de salida una señal sin ruido representativa del habla emitida por el portador de los casos de la Figura 7.

La Figura 9 ilustra dos cronogramas correspondientes, respectivamente, a un ejemplo de la señal ruidosa recogida por los micros, y de la señal recogida por un sensor fisiológico que permite distinguir los periodos de habla y los periodos de silencio del locutor.

La Figura 1 ilustra de forma esquemática, en forma de bloques, las diferentes funciones puestas en práctica por la invención.

El proceso de la invención se pone en práctica mediante medios software, esquematizados mediante un cierto número de bloques funcionales correspondientes a algoritmos apropiados ejecutados por un microcontrolador o un procesador digital de señal. Aunque, para clarificar la exposición, las diferentes funciones se representen bajo la forma de módulos diferenciados, dichas funciones aplican elementos comunes y se corresponden en la práctica con una pluralidad de funciones globalmente ejecutadas por un mismo ordenador.

La señal de la que se desea suprimir los ruidos es emitida por una red de sensores microfónicos la cual, en la configuración mínima ilustrada, puede ser simplemente una red de dos sensores dispuestos según una configuración predeterminada, estando cada sensor constituido por un micro respectivo correspondiente 10, 12.

La invención puede, no obstante, generalizarse a una red de más de dos sensores microfónicos, y/o a unos sensores microfónicos de los cuales cada sensor esté constituido por una estructura más compleja de la de un solo micro, por ejemplo, una combinación de varios micros y/o de otros sensores de habla.

Los micros 10, 12 son unos micros que captan la señal emitida por la fuente de señal útil (la señal de habla del locutor), y la diferencia de posición entre los dos micros determina un conjunto de desfases y variaciones de amplitud en el registro de las señales emitidas por la fuente de señal útil.

En la práctica, los dos micros 10, 12, son unos micros omnidireccionales dispuestos a unos centímetros uno de otro

sobre la luz cenital de un habitáculo de automóvil, sobre el frontal de una radio del automóvil o de un emplazamiento apropiado del salpicadero, o bien sobre la carcasa de uno de los auriculares de unos cascos de audio, etc.

Como se podrá apreciar, la técnica de la invención permite asegurar una supresión de ruido eficaz incluso para micros muy próximos, es decir separados entre ellos por una separación d tal que el retardo de fase máxima de una señal captada por un micro y después por el otro sea inferior al periodo de muestreo del convertidor de digitalización de las señales. Ello se corresponde con una distancia máxima d del orden de 4,7 cm para una frecuencia de muestreo F_e de 8 kHz (y una separación d media menor para una frecuencia doble, etc.).

Una señal de habla emitida por un locutor próximo alcanzará uno de los micros antes que el otro y ofrecerá, por tanto, un retardo y, por tanto, un desfase φ, sensiblemente constante. Respecto del ruido, puede sin duda existir igualmente un desfase entre los dos micros 10 y 12. Por contra, al estar unida la noción de desfase a la noción de onda incidente, se puede esperar que el desfase sea diferente al del habla. Por ejemplo, si un ruido directivo es dirigido en el sentido opuesto al de la boca, su desfase será de - φ si el desfase para la voz es φ. En el supuesto de la invención, la reducción de ruido sobre las señales captadas para los micros 10 y 12 no se opera en el dominio frecuencial (como sucede a menudo en las técnicas convencionales de supresión de ruido) sino en el dominio temporal.

Esta reducción de ruido se opera mediante un algoritmo que pretende la función de transferencia entre uno de los micros (por ejemplo el micro 10) y el otro micro (el micro 12) por medio de un combinador adaptativo 14 que emplea un filtro predictivo 16 de tipo LMS (*Least Mean Squares*, mínimos cuadrados medios). La salida del filtro 16 se sustrae en la convolución 18 de la señal del micro 10 para dar una señal S de ruido suprimido, aplicada otra vez al filtro 16 para permitir su adaptación iterativa en función del error de predicción. Es así posible predecir a partir de la señal captada por el micro 12 el componente de ruido contenido en la señal captada por el micro 10 (identificando la función de transferencia la transferencia del ruido).

La búsqueda adaptativa de la función de transferencia entre los dos micros no se opera más que durante las fases de ausencia de habla. Para ello, la adaptación iterativa del filtro 16 no se activa más que cuando un detector 20 de actividad vocal VAD (Detector de Actividad Vocal) accionado por un sensor 22 indica que el locutor próximo no está hablando. Esta función es esquematizada por el conmutador 24: en ausencia de la señal de habla manifestada por el detector de actividad vocalmente, el combinador adaptativo 14 pretende optimizar la función de transferencia entre los dos micros 10 y 12 para reducir el componente de ruido (posición cerrada del conmutador 24 como se ilustra en la figura); por contra en presencia de una señal de habla, manifestada por el detector de actividad vocal 20, el combinador adaptivo 14 congela los parámetros del filtro 16 en el valor en el que se encontraban justo antes de que se detecte el habla (apertura del conmutador 24), lo que evita cualquier degradación de la señal de habla del locutor próximo.

Se advertirá que esta manera de proceder no es molesta incluso en presencia de un entorno ruidoso evolutivo, pues las actualizaciones de los parámetros del filtro 16 son muy frecuentes ya que intervienen cada vez que el locutor próximo cesa de hablar.

Como característica distintiva de la invención, el filtrado del combinador adaptativo 14 es un filtrado de retardo fraccionario, es decir que permite aplicar un filtrado entre las señales captadas por los dos micros teniendo en cuenta un retardo inferior a la duración de una muestra de digitalización de las señales.

Es sabido que una señal temporal *x(t)* de paso banda [0,*Fe*/2] puede reconstruirse de manera perfecta a partir de la serie discreta *x(k)*, donde las muestras *x(k)* se corresponden con los valores de *x(t)* en los instantes *k.Te* siendo (*Te* = 1/*Fe*) el periodo de muestreo).

La expresión matemática es la siguiente:

$$x(t) = \sum_k x(k).\mathrm{sinc}\left(\frac{t - k.Te}{Te}\right)$$

Definiéndose la función seno cardinal mediante:

$$\mathrm{sinc}(t) = \frac{\mathrm{sen}(pi * t)}{pi * t}$$

La Figura 2 da una representación gráfica de esta función sinc *(t)*. Como se puede constatar decrece rápidamente, con la consecuencia de que un número finito y relativamente débil de coeficientes *k* en la suma produce una muy buena aproximación al resultado real.

Para una señal digitalizada con un periodo de muestreo *Te*, el intervalo o separación entre dos muestras corresponde de manera temporal a una duración de *Te* secundaria.

La serie *x(n)* de n muestras sucesivas digitalizadas de la señal captada puede así representarse mediante la expresión siguiente para todo n entero:

$$x(n.Te) = \sum_k x(k).\mathrm{sinc}\left(\frac{n.Te - k.Te}{Te}\right)$$

Se advertirá que en la suma del término en sinc es nulo para todo *k*, salvo para *k* = *n*.

La Figura 3a ofrece una representación gráfica de esta función.

Se quiere calcular esta misma serie *x(n)* desplazada un valor fraccionario τ es decir con un retardo inferior a la situación de una muestra de digitalización Te, la expresión anterior se convierte en:

$$x(n.Te - \tau) = \sum_k x(k).\mathrm{sinc}\left(\frac{(n-k).Te - \tau}{Te}\right)$$

La Figura 3b ofrece una representación gráfica de esta función, para un ejemplo de valor fraccionario τ = 0,5 (media muestra).

La serie x'(n) (desplazada τ) puede considerarse como la convolución de *x(n)* mediante un filtro no causal *G* de forma que:

$$x'(n) = G \otimes x(n)$$

Se trata, por tanto, de determinar una estimación $\hat{G}$ de un filtro óptimo *G* de forma que:

$$\hat{H} = \hat{G} \otimes \hat{F} \qquad \text{y} \qquad G(k) = \mathrm{sinc}(k + \tau / Te)\,,$$

siendo $\hat{H}$ la estimación de la transferencia de ruido entre los dos micros, incluyendo un retardo fraccionario, y

siendo *F* la estimación de la respuesta acústica del entorno.

Para la estimación del filtro de transferencia de ruido entre los dos micros, la estimación $\hat{H}$ corresponde a un filtro que minimiza un error:

$$e(n) = MicAvant(n) - \hat{H} * MicArrière(n)$$

siendo MicAvant(n) y MicArriére(n) los valores respectivos de las señales emitidas por los sensores microfónicos 10 y 12.

La característica de este filtro es que no es causal, es decir que se sirve de las muestras futuras. En la práctica, esto significa que se produce un retardo entre el retardo de tratamiento algorítmico. Como no es causal, puede modelar un retardo fraccionario y puede por tanto escribirse $\hat{H} = \hat{G} \otimes \hat{F}$. (En el caso clásico de un filtro causal se tendría $\hat{H} = \hat{F}$).

Concretamente, en el algoritmo, la estimación de $\hat{H}$ tiene lugar directamente, por la minimización del error e(n) anterior, sin que haya necesidad de estimar separadamente $\hat{G}$ y $\hat{F}$.

En el caso clásico causal (por ejemplo para un filtro de anulación de eco), el error e(n) que hay que minimizar se escribe, en forma desarrollada:

$$e(n) = MicAvant(n) - \sum_{k=0}^{L-1} \hat{H}(k).MicArrière(n-k)$$

Siendo L la longitud del filtro.

En el caso de la presente invención (filtro no causal) el error se convierte en:

$$e(n) = MicAvant(n) - \sum_{k=0}^{L-1} \hat{H}(k).MicArrière(n-k)$$

Se advertirá que la longitud del filtro se ha doblado, para tener en cuenta muestras futuras.

La predicción del filtro *H* ofrece un filtro de retardo fraccionario el cual, idealmente y en ausencia de habla, anula el

ruido del micro 10 al ser preferente el micro 12 (como se ha indicado más arriba, en periodo de habla el filtro es sin embargo congelado para evitar cualquier degradación del habla local).

Concretamente, el filtro $\hat{H}$ calculado para el algoritmo adaptativo que estima la transferencia de ruido entre el micro 10 y el micro 12, puede verse como la convolución $\hat{H} = \hat{G} \otimes \hat{F}$ de los filtros $\hat{G}$ y $\hat{F}$ donde:

- $\hat{G}$ corresponde a la parte fraccionaria (con la forma seno cardinal), y

- $\hat{F}$ corresponde a la transferencia acústica entre los dos micros, es decir a la parte “medioambiental” del sistema, representativa de la acústica del volumen en el que opera aquél.

La Figura 4 ilustra un ejemplo de respuesta acústica entre los dos micros, bajo la forma de una característica que ofrece una amplitud *A* en función de los coeficientes *k* del filtro *F*. Las diferentes reflexiones del sonido que pueden intervenir en función del entorno, por ejemplo en los cristales u otras paredes de un habitáculo de coche, crean unos picos visibles en esta característica de respuesta acústica.

La Figura 5 ilustra un ejemplo del resultado de la convolución $G \otimes F$ de los dos filtros *G* (respuesta en seno cardinal) y *F* (entorno de utilización), bajo la forma de una característica que ofrece la amplitud *A* en función de los coeficientes *k* del filtro convolucionado.

La estimación $\hat{H}$ puede calcularse mediante un algoritmo LMS iterativo que pretende minimizar el error $y(n) - \hat{H} \otimes x(n)$ para converger hacia el filtro óptimo.

Los algoritmos de tipo LMS - o NLMS (Normalized LMS) que son una versión normalizada del LMS - son unos algoritmos relativamente simples y poco exigentes en términos de recursos de cálculo. Se trata de algoritmos conocidos de por sí, descritos por ejemplo por:

[1] B. Widrow, Adaptative Filters, Aspect of Network and System Theory, R.E. Kalman and N. De Claris Eds., New York: Holt, Rinehart and Winston, pp. 563 - 587, 1970;

[2] B. Widrow et al., Adaptative Noise Cancelling: Principles and Applications, Proc. IEEEE, Vol. 63, No 12 pp. 1692 - 1716, dic. 1975.

[3] B. Widrow y S. Stearns, Adaptative Signal Processing, Prentice -Hall Signal Processing Series, Alan V. Oppenheim Series Editor, 1985.

Como se ha indicado más arriba, para que el tratamiento precedente sea posible, es necesario disponer de un detector de actividad local que permita discriminar entre las fases de ausencia de habla (o que la adaptación del filtro permita optimizar la evolución del ruido) y de presencia de habla (o que los parámetros del filtro sean congelados en su último valor encontrado).

Más exactamente, el detector de actividad vocal es aquí, de modo preferente, un detector “perfecto”, es decir que emite una señal binaria (ausencia vs. presencia de habla). Se distingue así de la mayor parte de los detectores de actividad vocal utilizados en los sistemas de supresión de ruido conocidos, que emiten solamente una probabilidad de ausencia de habla variable entre un 0 y un 100 % de forma continua o en pasos sucesivos. Con tales detectores basados solamente en una probabilidad de ausencia de habla, las falsas detecciones pueden ser importantes en los entornos ruidosos. Para ser “perfecto”, el detector de actividad vocal no se puede basar únicamente en la señal captada por los micros, sino que debe disponer de una información adicional que permita discriminar las fases de habla y de silencio del locutor próximo.

Un primer ejemplo de un detector de este tipo se ilustra mediante la Figura 6, donde el detector vocal 20 opera como respuesta a la señal producida por una cámara.

Esta cámara es, por ejemplo, una cámara 26 instalada en el habitáculo de un vehículo automóvil, y orientada de forma que su ángulo de campo 28 englobe en todas las circunstancias la cabeza del conductor 30, considerado como el locutor próximo. La señal emitida por la cámara 26 se analiza para determinar, de acuerdo con el movimiento de la boca y de los labios, si el locutor habla o no.

Con este fin, se pueden utilizar unos algoritmos de detección de la zona de la boca en una imagen de un rostro, y de seguimiento del movimiento de los labios (lip contour tracking) como los analizados en especial por:

[4] G. Potamianos et al., Audio-Visual Automatic Speech Recognition: An Overview, Audio-Visual Speech Processing, G. Bailly et al. Eds., MIT Press, pp. 1 - 30, 2004.

Este documento describe, con carácter general, el aporte de una información visual como complemento de una señal de audio para, en especial, efectuar el reconocimiento vocal en condiciones acústicas degradadas.

Los datos de vídeo, vienen así a añadirse a los datos de audio convencionales para mejorar la información vocal.

Este tratamiento podrá utilizarse en el marco de la presente invención para distinguir entre las fases de habla y las fases de silencio del locutor. Para tener en cuenta el hecho de que en un habitáculo de automóvil los movimientos del usuario son lentos mientras que los movimientos de la boca son rápidos, se puede, por ejemplo, una vez localizada la boca, comparar dos imágenes consecutivas y evaluar la separación en un mismo píxel.

La ventaja de esta técnica de análisis de imagen consiste en disponer de una información complementaria totalmente independiente del entorno de ruido acústico.

Otro ejemplo de sensor utilizable para la detección vocal “perfecta” es un sensor fisiológico susceptible de detectar ciertas vibraciones locales del locutor que no estén o que estén escasamente corrompidas por el ruido circundante.

Un sensor de este tipo puede estar, en especial, constituido por un acelerómetro o por un sensor piezoeléctrico aplicado a la mejilla o la sien del locutor. En efecto, cuando una persona emite un sonido vocalizado (es decir un compuesto de habla cuya emisión se acompaña con una vibración de las cuerdas vocales), se propaga una vibración desde las cuerdas vocales hasta la faringe y en la cavidad buconasal, donde es modulada, amplificada y articulada. La boca, el velo del paladar, la faringe, los senos y las fosas nasales sirven luego de caja de resonancia a este sonido vocalizado y, siendo su pared elástica, vibran a su vez y estas vibraciones son transmitidas mediante conducción ósea interna y son perceptibles al nivel de la mejilla y de la sien.

Estas vibraciones al nivel de la mejilla y de la sien presentan la característica de estar, por naturaleza, muy poco corrompidas por el ruido ambiente: en efecto, en presencia de ruidos exteriores, incluso importantes, los tejidos de la mejilla y de la sien apenas vibran, y ello cualquiera que sea la composición espectral del ruido exterior.

Un sensor fisiológico que recoja estas vibraciones vocales desprovistas de ruido proporciona una señal representativa de la presencia o de la ausencia de los sonidos vocalizados emitidos por el locutor, que permiten, por tanto, discriminar muy bien las fases de habla y las fases de silencio del locutor.

Un sensor fisiológico de este tipo puede, en particular, incorporarse a un conjunto combinado de micro/cascos como por ejemplo el ilustrado en la Figura 7.

En esta figura, la referencia 32 designa globalmente los cascos según la invención, los cuales incorporan dos auriculares 34 unidos por un arco. Cada uno de los auriculares está, de modo preferente, constituido por una carcasa cerrada 36, que aloja un transductor de reproducción sonora, aplicado alrededor de la oreja del usuario con la interposición de una almohadilla 38 que aísla el oído del exterior.

El sensor fisiológico 40 que sirve para la detección de actividad vocal es, por ejemplo, un acelerómetro integrado en la almohadilla 38 para acoplarse a la mejilla o la sien del usuario con un acoplamiento lo más estrecho posible. El sensor fisiológico 40 puede, en particular, situarse sobre la cara interior de la piel de la almohadilla 38 de manera que, una vez que los cascos se colocan en posición, el sensor se aplique contra la mejilla o la sien del usuario por efecto de una ligera presión derivada del aplastamiento del material de la almohadilla, solamente con la interposición de la piel exterior de esta almohadilla.

Los cascos incorporan igualmente los micros 10, 12 del circuito de recogida y de supresión de ruido del habla del locutor. Estos dos micros son unos micros omnidireccionales situados sobre la carcasa 36, y están dispuestos con el micro 10 situado en posición adelantada (más cerca de la boca del portador de los cascos) y el micro 12 situado más hacia atrás. Por otro lado, la posición de alineamiento 42 de los dos micros 10, 12 está aproximadamente dirigida hacia la boca 44 del portador de los casos.

La Figura 8 es un esquema de bloques que muestra las diferentes funciones empleadas por el conjunto micro/cascos de la Figura 7.

Se vuelven a encontrar en este figura los dos micros 10 y 12, así como el detector de actividad vocal 20. El micro delantero 10 es el micro principal y el micro trasero 12 sirve de entrada al filtro adaptativo 16 del combinador 14. El detector de actividad vocal 20 es controlado por la señal emitida por el sensor fisiológico 40 con, por ejemplo, el alisado de la potencia de la señal emitida por el sensor 40.

$$\text{Puissance}_{\text{capteur}}(n) = \alpha.\text{puissance}_{\text{capteur}}(n-1) + (1-\alpha), (\text{capteur}(n))^2$$

siendo α una constante de alisado próxima a 1. Basta entonces con fijar un umbral ζ de forma que el umbral sea sobrepasado en el momento en que el locutor habla.

La Figura 9 ilustra la traza de las señales recogidas:

- la señal S_{10} del cronograma superior se corresponde con la que es captada por el micro delantero 10: se ve que es imposible producir a partir de esta señal (ruidosa) una discriminación eficaz entre las fases de presencia y ausencia de habla.

- la señal S_{40} del cronograma de abajo se corresponde con el que emite simultáneamente el sensor fisiológico 40: las fases sucesivas de ausencia y presencia de habla están marcadas de manera muy visible. La señal

binaria designada como VAD se corresponde con la indicación emitida por el detector de actividad vocal 20 ('1' = presencia de habla; '0' = ausencia de habla), después de la evaluación de la potencia de la señal S_{40} y la comparación en relación al umbral ζ predefinido.

La señal emitida por el sensor fisiológico 40 puede utilizarse no solamente como señal de entrada de un detector de actividad vocal, sino igualmente para enriquecer la señal captada por los micros 10 y 12, en especial en el registro bajo del espectro.

Por supuesto, las señales emitidas por el sensor fisiológico, que se corresponden a los sonidos vocalizados, no son estrictamente hablando habla, puesto que el habla no está solamente formada por los sonidos vocalizados, sino que contiene componentes que no nacen al nivel de las cuerdas vocales: el contenido frecuencial es por ejemplo mucho más rico con el sonido proveniente de la garganta y emitido por la boca. Además, la conducción ósea interna y el atravesar la piel tienen como efecto filtrar determinados componentes vocales.

Por otro lado, en razón del filtrado debido a la propagación de las vibraciones hasta la sien o la mejilla, la señal recogida por el sensor fisiológico se utiliza típicamente en bajas frecuencias, principalmente en la región inferior del espectro sonoro (típicamente de 0 a 1500 Hz).

Pero como los ruidos que generalmente se perciben en un entorno habitual (calle, metro, tren, …) están mayoritariamente concentrados en frecuencias bajas, la señal de un sensor fisiológico presenta la ventaja considerable de estar naturalmente desprovista de componentes parásitos de ruido y será, por tanto, posible utilizar esta señal en el registro bajo del espectro, completándolo en el registro alto del espectro (por encima de 1500 Hz) mediante las señales (ruidosas) recogidas por los micros 10 y 12, después de haber sometido estas señales a una reducción de ruido operada por el combinador adaptativo 14.

El espectro completo es reconstruido por el bloque de mezcla 46 que recibe paralelamente: la señal del sensor fisiológico 40 para el registro bajo del espectro, y la señal de los micros 10 y 12 después de la supresión del ruido mediante el combinador adaptativo 14 para el registro alto del espectro. Esta reconstrucción se produce mediante la suma de las señales, que son aplicadas en sincronía al bloque de mezcla 46 para evitar cualquier deformación.

La señal resultante emitida por el bloque 46 puede someterse a una reducción final de ruido por el circuito 48, operada en el dominio frecuencial según una técnica convencional comparable a la descrita, por ejemplo, en el documento WO 2007/099222 A1 (Parrot), para producir como salida la señal últimas desprovista de ruido.

El establecimiento de esta técnica resulta, sin embargo, fuertemente simplificado con respecto a la que se da a conocer por ejemplo en el documento precitado. En efecto, en el caso presente, ya no es necesario evaluar una probabilidad de presencia de habla a partir de la señal recogida puesto que esta información puede obtenerse obtenida directamente por el bloque de detección de actividad 20 en respuesta a la emisión de la detección de sonido vocalizado detectada por el sensor fisiológico 40. El algoritmo puede así simplificarse haciéndolo más eficaz y rápido.

La reducción de ruido frecuencial se opera, de modo ventajoso, de manera diferente en presencia y en ausencia de habla (información dada por el detector de actividad vocal perfecto 20):

- en ausencia de habla, la reducción de ruido es máxima en todas la bandas de frecuencias, es decir que la ganancia correspondiente a la supresión de ruido máxima es aplicada de la misma manera sobre todos los componentes de la señal (puesto que existe la seguridad en este caso de que aquella no contiene componente útil);
- por el contrario, en presencia de habla, la reducción de ruido es una reducción frecuencial, aplicada de manera diferenciada sobre cada banda de frecuencias según el esquema clásico.

El sistema que se acaba de describir permite obtener excelentes rendimientos globales, típicamente del orden de 30 a 40 dB de reducción de ruido sobre la señal de habla del locutor próximo. El combinador adaptativo 14 que opera sobre las señales captadas por los micros 10 y 12 permite, en particular, con el filtrado de retardo fraccionario, obtener muy buenos rendimientos de supresión de ruido en frecuencias altas.

Gracias a la eliminación de todos los ruidos parásitos, ello da la impresión al locutor distante (con el que el portador de los cascos está en comunicación) que su interlocutor (el portador de los cascos) se encuentre en una habitación silenciosa.

REIVINDICACIONES

1. Un equipo de audio, que comprende:

- un conjunto de dos sensores microfónicos (10, 12) aptos para recoger el habla del usuario del equipo y para emitir unas señales de habla ruidosas respectivas;

- medios de muestreo de las señales de habla emitidas por los sensores microfónicos; y

- medios de supresión de ruido de una señal de habla, que reciben como entrada las muestras de las señales de habla emitidas por los dos sensores microfónicos, y emiten como salida una señal de habla carente de ruido representativa del habla emitida por el usuario del equipo,

en el que los medios de supresión de ruido son medios de reducción de ruido no frecuencial que comprenden un combinador de filtro adaptativo (14) de las señales emitidas por los dos sensores microfónicos, que operan mediante búsqueda iterativa con el objeto de anular el ruido captado por uno de los sensores microfónicos (10) en base a una referencia de ruido dada por la señal emitida por el otro sensor microfónico (12);

estando el equipo **caracterizado por que**:

- el filtro adaptativo (16) es un filtro de retardo fraccionario, apto para modelar un retardo inferior al periodo de muestreo de los medios de muestreo;

- el equipo comprende además medios de detección de actividad vocal (20, 22) aptos para emitir una señal representativa de la presencia o de la ausencia de habla por el usuario del equipo; y

- el filtro adaptativo recibe igualmente como entrada la señal de presencia o de ausencia de habla para, de manera selectiva: i) o bien operar una búsqueda adaptativa de los parámetros del filtro en ausencia de habla, ii) o bien congelar estos parámetros del filtro en presencia de habla.

2. El equipo de audio según la reivindicación 1, en el que el filtro adaptativo (16) es apto para estimar un filtro óptimo H de forma que:

$$\hat{H} = \hat{G} \otimes \hat{F}$$

con:

$$x'(n) = G \otimes x(n) \quad \textbf{y} \quad G(k) = \mathrm{sinc}(k + \tau / Te),$$

representando	$\hat{H}$	la estimación del filtro óptimo H, la transferencia de ruido entre los dos sensores microfónicos para una respuesta de impulso que incluye un retardo fraccionario,
representando	$\hat{G}$	la estimación del filtro del retardo fraccionario G entre los dos sensores microfónicos,
representando	$\hat{F}$	la estimación de la respuesta acústica del entorno,
indicando	$\otimes$	una convolución,
siendo	$x(n)$	la serie de muestras de la señal de entrada del filtro H,
siendo	$x'(n)$	la serie x(n) desplazada el retardo τ,
siendo	Te	el periodo de muestreo de la señal de entrada del filtro H,
siendo	τ	dicho retardo fraccionario, igual a un submúltiplo de Te, e
indicando	sinc	la función seno cardinal.

3. El equipo de audio según la reivindicación 1, en el que el filtro adaptativo es un filtro de logaritmo de predicción lineal de tipo mínimos cuadrados medios, LMS.

4. El equipo de audio según la reivindicación 1, en el que:

- el equipo comprende además una cámara de vídeo (26) dirigida hacia el usuario (30) del equipo y apta para captar una imagen de éste, y

- los medios de detección de actividad vocal (20) comprenden medios de análisis de vídeo aptos para analizar la imagen producida por la cámara y emitir como respuesta dicha señal de presencia o de ausencia de habla

por dicho usuario.

5. El equipo de audio según la reivindicación 1, en el que:

- el equipo comprende además un sensor fisiológico (40) apto para situarse en contacto con la cabeza del usuario del equipo para quedar allí acoplada con el fin de captar las vibraciones vocales acústicas transmitidas por conducción ósea interna, y

- los medios de detección de actividad vocal (20) comprenden unos medios aptos para analizar la señal emitida por el sensor fisiológico y para emitir como respuesta dicha señal de presencia o de ausencia de habla por dicho usuario.

6. El equipo de audio según la reivindicación 5, en el que los medios de detección de actividad vocal comprenden medios de evaluación de la energía de la señal emitida por el sensor fisiológico, y unos medios de umbral.

7. El equipo de audio según la reivindicación 6, en el que el equipo consiste en unos cascos de audio del tipo combinado micro/cascos, que comprende:

- unos auriculares (34) que incorporan cada uno un transductor de reproducción sonora de una señal de audio alojada en una carcasa (36) provista de una almohadilla (38) circumaural;

- dichos dos sensores microfónicos (10, 12) dispuestos sobre la carcasa de uno de los auriculares; y

- dicho sensor fisiológico (40), incorporado en la almohadilla de dichos auriculares y colocado en una región de ésta apta para situarse en contacto con la mejilla o con la sien del portador de los casos.

8. El equipo de audio según la reivindicación 7, en el que los dos sensores microfónicos (10, 12) están alineados en una red lineal siguiendo una dirección principal (42) dirigida hacia la boca (44) del usuario del equipo.

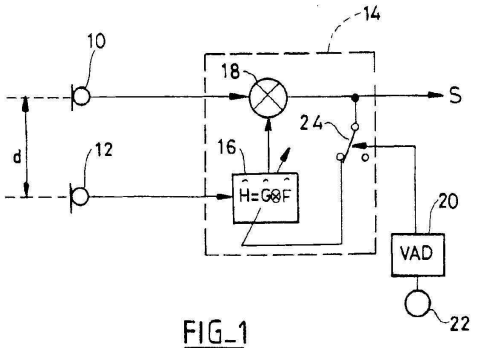

FIG_1

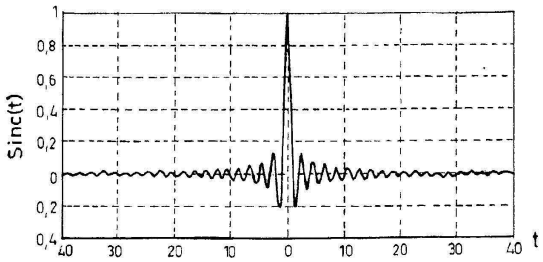

FIG_2

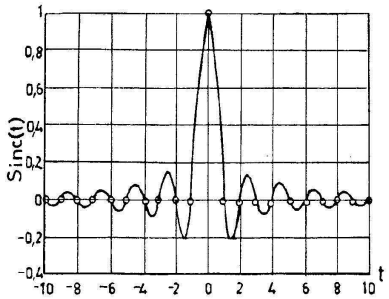

FIG_3a

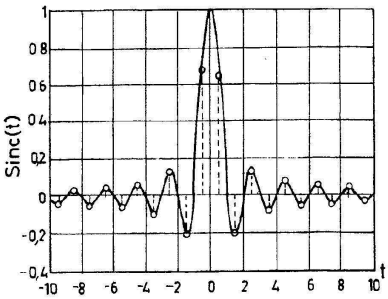

FIG_3b

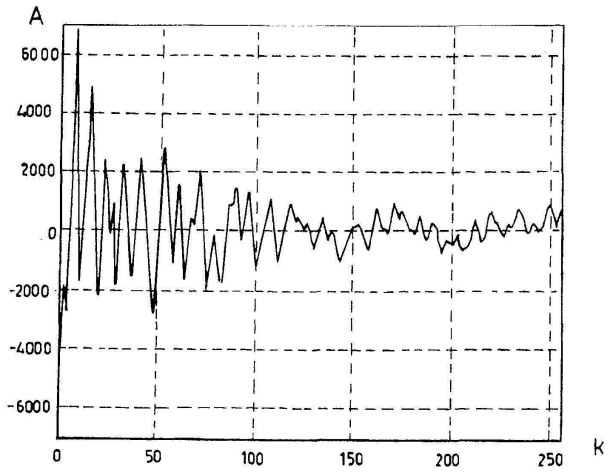
A
6000
4000
2000
0
-2000
-4000
-6000
0
50
100
150
200
250
k

FIG_4

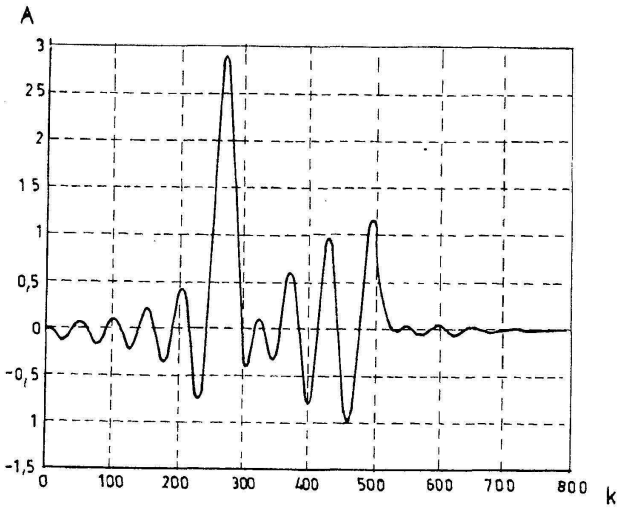
A
3
2,5
2
1,5
1
0,5
0
-0,5
1
-1,5
0
100
200
300
400
500
600
700
800
k

FIG_5

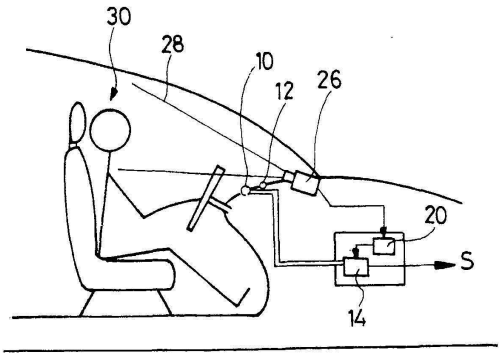

FIG_6

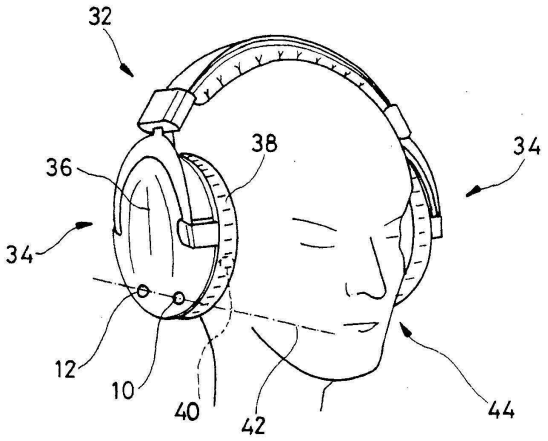

FIG_7

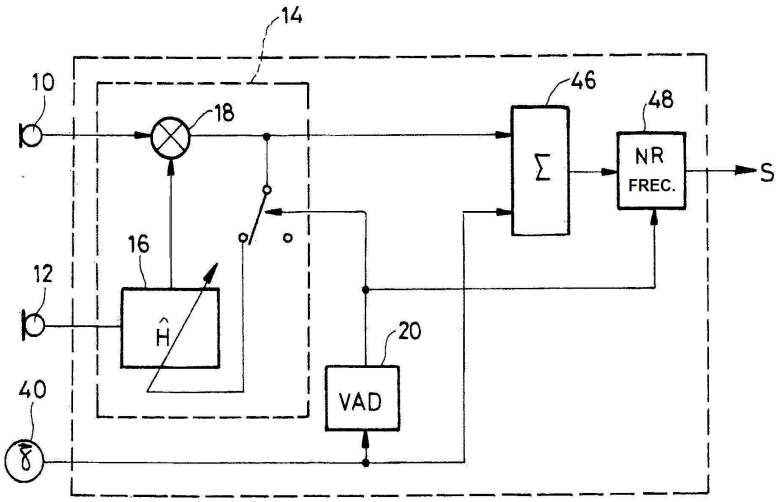

FIG_8

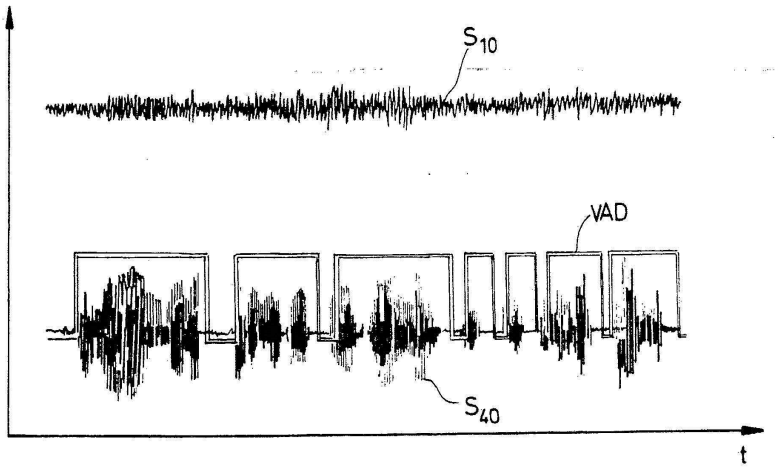

FIG_9