(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 400 700**

(51) Int. CI.:

G06K 9/00 (2006.01)
G10L 15/04 (2006.01)
G10L 17/00 (2006.01)
G10L 21/04 (2006.01)
H04N 5/60 (2006.01)
G10L 11/00 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **26.02.2002 E 02721201 (8)**

(97) Fecha y número de publicación de la concesión europea: **28.11.2012 EP 1393300**

(54) Título: **Segmentación de señales de audio en eventos auditivos**

(30) Prioridad:

25.05.2001 US 293825 P
11.01.2002 US 45644
23.01.2002 US 351498 P
12.02.2002 WO PCT/US02/04317

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
11.04.2013

(73) Titular/es:

DOLBY LABORATORIES LICENSING CORPORATION (100.0%)
100 POTRERO AVENUE
SAN FRANCISCO, CALIFORNIA 94103-4813, US

(72) Inventor/es:

CROCKETT, BRETT, G.

(74) Agente/Representante:

LINAGE GONZÁLEZ, Rafael

DESCRIPCIÓN

Segmentación de señales de audio en eventos auditivos

Campo técnico

La presente invención pertenece al campo del procesamiento psicoacústico de señales de audio. En particular, la invención se refiere a aspectos de la división o segmentación de señales de audio en “eventos auditivos”, cada uno de los cuales tiende a ser percibido como separado y distinto, y a aspectos de la generación de representaciones de información reducida de señales de audio en base a eventos auditivos y, opcionalmente, también en base a las características o rasgos de señales de audio dentro de tales eventos auditivos. Los eventos auditivos pueden ser útiles como definen los “Segmentos de Audio” MPEG-7 como se propone por la “ISO/IEC.ITC 1/SC 29/WG 11.”

Antecedentes de la técnica

La división de sonidos en unidades o segmentos percibidos como separados y distintos se denomina “análisis de eventos auditivos” o “análisis de escenas auditivas” (ASA, del inglés “*Auditory Scene Analysis*”). Se establece una amplia discusión del análisis de escenas auditivas por Albert S. Bregman en su libro *Auditory Scene Analysis - The Perceptual Organization of Sound,* (Massachusetts Institute of Technology, 1991, cuarta edición, 2001, Second MIT Press paperback edition). Además, la patente de Estados Unidos 6.002.776 de Bhadkamkar, y otros, 14 de diciembre de 1999, cita publicaciones que datan de 1979 como “trabajo de la técnica anterior relacionado con la separación del sonido por el análisis de escenas auditivas”. Sin embargo, la patente de Bhadkamkar, y otros, desincentiva el uso práctico del análisis de escenas auditivas, concluyendo que “las técnicas que impliquen análisis de escenas auditivas, aunque interesantes desde un punto de vista científico como modelos del procesamiento auditivo humano, son actualmente demasiado especializadas y exigentes computacionalmente para ser consideradas técnicas prácticas para separar el sonido hasta que se haga un progreso fundamental”.

Hay muchos métodos diferentes para extraer características o rasgos del audio. Siempre que las características o rasgos estén adecuadamente definidos, su extracción se puede realizar usando procesos automatizados. Por ejemplo, la “ISO/IEC.ITC 1/SC 29/WG 11” (MPEG) está estandarizando actualmente una variedad de descriptores de audio como parte del estándar MPEG-7. Una deficiencia común de tales métodos es que ignoran el análisis de escenas auditivas. Tales métodos buscan medir, periódicamente, ciertos parámetros de procesamiento de señales “clásicos” tales como tono, amplitud, potencia, estructura armónica y planicidad espectral. Tales parámetros, aunque proporcionan información útil, no analizan y caracterizan señales de audio en elementos percibidos como separados y distintos de acuerdo con la cognición humana. Sin embargo, los descriptores MPEG-7 pueden ser útiles para caracterizar un Evento Auditivo identificado de acuerdo con aspectos de la presente invención.

El documento “Sound onset detection by applying psychoacoustic knowledge" de A. Klapuri (ICASSP, IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING - PROCEEDINGS 1999 IEEE, vol. 6, 15 de marzo de 1999, páginas 3089-3092, XP010328057, DOI: DOI: 10.1109/ICASSP.199.757494, ISBN: 978-0-7803-5041-0) divulga un sistema para detectar comienzos perceptuales de sonidos en señales acústicas, donde el sistema determina principios de sonido que tienen imperfecciones de comienzo y utiliza procesamiento en modo banda y un modelo psicoacústico de codificación de intensidad para combinar resultados de varias bandas de frecuencia.

El documento “Tempo and beat analysis of acoustic signals" de E. Scheirer (THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, AMERICAN INSTITUTE OF PHYSICS FOR THE ACOUSTICAL SOCIETY OF AMERICA, NUEVA YORK, NY, EE.UU., vol. 103, nº 1, 1 de enero de 1998, páginas 588-601, XP012000051, ISSN: 0001-4966, DOI: DOI: 10.1121/1.421129) divulga un método para usar un número reducido de filtros de paso de banda y bancos de filtros de peine paralelos para analizar el tempo y extraer el compás de señales musicales.

El documento “Computer modeling of sound for transformation and synthesis of musical signals" de P. Masri (Thesis 1 de diciembre de 1996, véase www.mp3-tech.org/programmer/docs/Masri_thesis.pdf), capítulo cinco "improved synthesis of attack transients", p. 125 - 147, divulga métodos para la detección y ubicación de eventos transitorios en base a la distribución de energía (véase la sub-sección 5.2.1), la envolvente de ataque (véase la sub-sección 5.2.2), y la disimilitud espectral (véase la sub-sección 5.2.3).

Sumario de la invención

De acuerdo con la presente invención, se proporciona un método para dividir cada uno de los múltiples canales de señales de audio digital en eventos auditivos de acuerdo con la reivindicación 1. Reivindicaciones dependientes se refieren a realizaciones preferidas de la presente invención.

Según aspectos de la presente invención, se proporciona un proceso eficiente computacionalmente para dividir audio en segmentos temporales o “eventos auditivos” que tienden a ser percibidos como separados o distintos. Las ubicaciones de los límites de estos eventos auditivos (dónde comienzan y finalizan con respecto al tiempo)

proporcionan información valiosa que se puede utilizar para describir una señal de audio. Las ubicaciones de los límites de un evento auditivo se pueden ensamblar para generar una representación de información reducida, "firma" o "huella dactilar" de una señal de audio que pueda ser almacenada para uso, por ejemplo, en análisis comparativos con otras firmas generadas similarmente (como, por ejemplo, en una base de datos de trabajos conocidos).

Bregman observa que "oímos unidades discretas cuando el sonido cambia abruptamente de timbre, tono, volumen, o (en menor medida) ubicación en el espacio" (*Auditory Scene Analysis - The Perceptual Organization of Sound,* arriba en página 469). Bregman también discute la percepción de corrientes de sonido múltiples y simultáneas cuando, por ejemplo, están separadas en frecuencia.

Con el fin de detectar cambios en timbre y tono y ciertos cambios en amplitud, el proceso de detección de eventos de audio según un aspecto de la presente invención detecta cambios en la composición espectral con respecto al tiempo. Cuando se aplica a una disposición de sonido multicanal en la que los canales representan direcciones en el espacio, el proceso según un aspecto de la presente invención también detecta eventos auditivos que resultan de cambios en la ubicación espacial con respecto al tiempo. Opcionalmente, según otro aspecto de la presente invención, el proceso también puede detectar cambios en amplitud con respecto al tiempo que no serían detectados detectando cambios en la composición espectral con respecto al tiempo.

En su implementación menos exigente computacionalmente, el proceso divide el audio en segmentos de tiempo analizando toda la banda de frecuencia (audio con ancho de banda completo) o sustancialmente toda la banda de frecuencia (en implementaciones prácticas, se emplea a menudo un filtrado de limitación de banda en los extremos del espectro) y dando el mayor peso a las componentes de señales de audio más fuertes. Este enfoque aprovecha un fenómeno psicoacústico en el cual, en escalas de tiempo más pequeñas (20 milisegundos (ms) y menos), el oído puede tender a enfocarse en un único evento auditivo en un tiempo dado. Esto implica que, aunque pueden estar sucediendo múltiples eventos auditivos al mismo tiempo, una componente tiende a ser perceptualmente más prominente y se puede procesar individualmente como si fuera el único evento que estuviera teniendo lugar. Aprovechar este efecto también permite que la detección del evento auditivo se escale con la complejidad del audio que se está procesando. Por ejemplo, si la señal de audio de entrada que se está procesando es un solo de un instrumento, los eventos auditivos que se identifican serán probablemente las notas individuales que se están tocando. Del mismo modo para una señal de voz de entrada, las componentes individuales del discurso, las vocales y consonantes por ejemplo, serán identificadas probablemente como elementos de audio individuales. Según aumenta la complejidad del audio, tal como música con un toque de tambor o múltiples instrumentos y voz, la detección del evento auditivo identifica el elemento de audio "más prominente" (es decir, el más fuerte) en cualquier momento dado. Alternativamente, el elemento de audio más prominente puede ser determinado teniendo en consideración el umbral de audición y la respuesta de frecuencia.

Aunque las ubicaciones de los límites del evento auditivo computados a partir de audio de ancho de banda completo proporcionan información útil relacionada con el contenido de una señal de audio, se podría desear proporcionar información adicional que describa adicionalmente el contenido de un evento auditivo para uso en análisis de señales de audio. Por ejemplo, una señal de audio podría ser analizada a través de dos o más sub-bandas de frecuencia, y la ubicación de eventos auditivos de sub-banda de frecuencia determinada y usada para transportar información más detallada sobre la naturaleza del contenido de un evento auditivo. Tal información detallada podría proporcionar información adicional no disponible del análisis de banda ancha.

Así, opcionalmente, de acuerdo con aspectos adicionales de la presente invención, a costa de una mayor complejidad computacional, el proceso también puede tomar en consideración cambios en la composición espectral con respecto al tiempo en sub-bandas discretas de frecuencia (fijas o determinadas dinámicamente o sub-bandas tanto fijas como determinadas dinámicamente) en lugar de la anchura de banda completa. Este enfoque alternativo tendría en cuenta más de una corriente de audio en diferentes sub-bandas de frecuencia en lugar de asumir que una única corriente es perceptible en un momento en particular.

Incluso un proceso simple y computacionalmente eficiente de acuerdo con aspectos de la presente invención ha resultado útil para identificar eventos auditivos.

Un proceso de detección de eventos auditivos de acuerdo con la presente invención puede implementarse dividiendo una forma de onda de audio de dominio tiempo en intervalos de tiempo o bloques y convirtiendo entonces los datos de cada bloque en el dominio frecuencia, usando o bien un banco de filtros o bien una transformación tiempo-frecuencia, como la FFT. La amplitud del contenido espectral de cada bloque se puede normalizar con el fin de eliminar o reducir el efecto de los cambios de amplitud. Cada representación resultante de dominio frecuencia proporciona una indicación del contenido espectral (amplitud en función de la frecuencia) del audio en el bloque particular. El contenido espectral de bloques sucesivos se compara, y se asume que los cambios mayores que un umbral indican el inicio temporal o fin temporal de un evento auditivo. La figura 1 muestra una forma de onda idealizada de un canal único de música orquestal que ilustra eventos auditivos. Los cambios espectrales que suceden cuando se toca una nueva nota desencadenan los nuevos eventos auditivos 2 y 3 en las muestras 2048 y 2560 respectivamente.

Como se mencionó anteriormente, con el fin de minimizar la complejidad computacional, sólo se puede procesar una única banda de frecuencias de la forma de onda de audio de dominio tiempo, preferiblemente o bien toda la banda de frecuencia del espectro (que puede ser alrededor de 50 Hz a 15 kHz en el caso de un sistema de música de calidad media) o sustancialmente toda la banda de frecuencia (por ejemplo, un filtro que define una banda puede excluir los extremos de frecuencia altos y bajos).

Preferiblemente, los datos del dominio frecuencia se normalizan, como se describe más abajo. El grado en el cual los datos de dominio frecuencia necesitan normalizarse da una indicación de amplitud. Por lo tanto, si un cambio en este grado supera un umbral predeterminado, esto también se puede asumir que indica un límite del evento. Los puntos de inicio y fin de evento resultantes de los cambios espectrales y de los cambios de amplitud se pueden ORed juntos de modo que se identifican los límites de evento resultantes de cualquier tipo de cambio.

En el caso de múltiples canales de audio, que representan cada uno una dirección en el espacio, cada canal se puede tratar de forma independiente y los límites de evento resultantes para todos los canales pueden ser entonces ORed juntos. Así, por ejemplo, un evento auditivo que cambia abruptamente de dirección probablemente resultará en un límite “fin de evento” en un canal y un límite “inicio de evento” en otro canal. Cuando son ORed juntos, se identificarán dos eventos. Así, el proceso de detección de evento auditivo de la presente invención es capaz de detectar eventos auditivos en base a cambios espectrales (timbre y tono), de amplitud y direccionales.

Como se ha mencionado anteriormente, como otra opción, pero a costa de una gran complejidad computacional, en lugar de procesar el contenido espectral de la forma de onda de dominio tiempo en una única banda de frecuencias, el espectro de la forma de onda de dominio tiempo anterior a la conversión de dominio frecuencia se puede dividir en dos o más bandas de frecuencia. Cada una de las bandas de frecuencia se puede convertir entonces al dominio frecuencia y procesar como si fuera un canal independiente de la forma descrita anteriormente. Los límites de evento resultantes se pueden entonces ORed juntos para definir los límites de evento para ese canal. Las múltiples bandas de frecuencia pueden ser fijas, adaptativas, o una combinación de fijas y adaptativas. Técnicas de filtro de rastreo empleadas en la reducción de ruido de audio y otras técnicas, por ejemplo, se pueden emplear para definir bandas de frecuencia adaptativas (por ejemplo ondas sinusoidales simultáneas dominantes a 800 Hz y 2 kHz podrían resultar en dos bandas adaptativamente determinadas, centradas en esas dos frecuencias). Aunque filtrar los datos antes de la conversión al dominio frecuencia es factible, es más óptimo que el audio de ancho de banda completo se convierta al dominio frecuencia y entonces sólo se procesan componentes de interés de sub-bandas de frecuencia. En el caso de convertir el audio de todo el ancho de banda usando la FFT, sólo se procesarían juntos los sub-contenedores correspondientes a sub-bandas de frecuencia de interés.

Alternativamente, en el caso de sub-bandas múltiples o canales múltiples, en lugar de Oring juntos límites de evento auditivo, lo que da como resultado alguna pérdida de información, se puede preservar la información de límite de evento.

Como se muestra en la figura 2, la magnitud del dominio frecuencia de una señal de audio digital contiene información útil de frecuencia a una frecuencia de Fs/2 donde Fs es la frecuencia de muestreo de la señal de audio digital. Dividiendo el espectro de la frecuencia de la señal de audio en dos o más sub-bandas (no necesariamente del mismo ancho de banda y no necesariamente hasta una frecuencia de Fs/2 Hz), las sub-bandas de frecuencia pueden analizarse a lo largo del tiempo de un modo similar a un método de detección de eventos auditivos de ancho de banda completo.

La información del evento auditivo de sub-banda proporciona información adicional sobre una señal de audio que describe más exactamente la señal y la diferencia de otras señales de audio. Esta capacidad diferenciadora mejorada puede ser útil si la información de firma de audio se va a utilizar para identificar señales de audio coincidentes de entre un gran número de firmas de audio. Por ejemplo, como se muestra en la figura 2, un análisis de eventos auditivos de sub-banda de frecuencia (con una resolución de límite de evento auditivo de 512 muestras) ha encontrado múltiples eventos auditivos de sub-bandas que comienzan, diversamente, en las muestras 1024 y 1536 y finalizan, diversamente, en las muestras 2560, 3072 y 3584. Es improbable que este nivel de detalle de señal resulte disponible de un único análisis de escenas auditivas de banda ancha.

La información de evento auditivo de su-banda se puede utilizar para obtener una firma de evento auditivo para cada sub-banda. Aunque esto aumentaría el tamaño de la firma de la señal de audio y posiblemente aumentaría el tiempo de computación requerido para comparar múltiples firmas, también podría reducir extremadamente la probabilidad de clasificar falsamente dos firmas como si fueran la misma. Se podría realizar una compensación entre el tamaño de la firma, complejidad computacional y precisión de la señal dependiendo de la aplicación. Alternativamente, en lugar de proporcionar una firma para cada sub-banda, los eventos auditivos se pueden ORed juntos para proporcionar un único conjunto de límites “combinados” de eventos auditivos (en muestras 1024, 1536, 2560, 3072 y 3584). Aunque esto se traduciría en una pérdida de información, proporciona un conjunto único de límites de eventos, que representan eventos auditivos combinados, que proporcionan más información que la información de una única sub-banda o análisis de banda ancha.

Aunque la información de evento auditivo de sub-banda de frecuencia por si sola proporciona información de señal

útil, la relación entre las ubicaciones de eventos auditivos de sub-banda de puede analizar y usar para proporcionar más percepción de la naturaleza de una señal de audio. Por ejemplo, la ubicación y fuerza de los eventos auditivos de sub-banda se pueden utilizar como una indicación de timbre (contenido de frecuencia) de la señal de audio. Los eventos auditivos que aparecen en sub-bandas y están relacionados armónicamente unos con otros también proporcionarían percepción útil con respecto a la naturaleza armónica del audio. La presencia de eventos auditivos en una única sub-banda también puede proporcionar información en cuanto a la naturaleza a modo de tono de una señal de audio. Analizar la relación de eventos auditivos de sub-banda de frecuencia a través de múltiples canales también puede proporcionar información de contenido espacial.

En el caso de analizar múltiples canales de audio, cada canal se analiza independientemente y la información de límite de evento auditivo de cada uno se puede retener por separado o combinar para proporcionar información de evento auditivo combinada. Esto es algo análogo al caso de múltiples sub-bandas. Eventos auditivos combinados pueden entenderse mejor por referencia a la figura 3 que muestra los resultados del análisis de escenas auditivas para una señal de audio de dos canales. La figura 3 muestra segmentos simultáneos en el tiempo de datos de audio en dos canales. El procesamiento ASA del audio en un primer canal, la forma de onda superior de la figura 3, identifica límites de evento auditivo en muestras que son múltiplos del tamaño de bloque de perfil espectral de la muestra 512, muestras 1024 y 1536 en este ejemplo. La forma de onda inferior de la figura 3 es un segundo canal y los resultados del procesamiento ASA en límites de evento de muestras que también son múltiplos del tamaño de bloque de perfil espectral, en muestras 1024, 2048 y 3072 en este ejemplo. Un análisis combinado de evento auditivo para ambos canales da como resultado segmentos de evento auditivo combinado con límites en las muestras 1024, 1536, 2084 y 3072 (los límites de evento auditivo de los canales se “ORed” juntos). Se apreciará que en la práctica la precisión de los límites de evento auditivo depende del tamaño del bloque de perfil espectral (N es 512 muestras en este ejemplo) porque los límites de evento sólo puedan ocurrir en los límites de bloque. Sin embargo, se ha encontrado que un tamaño de bloque de 512 muestras determina límites de evento auditivo con suficiente precisión para proporcionar resultados satisfactorios.

La figura 3A muestra tres eventos auditivos. Estos eventos incluyen la (1) parte tranquila de audio anterior al transitorio, (2) el evento transitorio, y (3) la porción sostenido / el eco del transitorio de audio. Una señal de voz se representa en la figura 3B con un evento de carácter sibilante predominantemente de frecuencia alta, y eventos mientras la sibilancia evoluciona o “morfea” en la vocal, la primera mitad de la vocal, y la segunda mitad de la vocal.

La figura 3 también muestra los límites de evento combinado cuando los datos del evento auditivo se comparten entre los bloques de datos de tiempo simultáneos de dos canales. Tal segmentación de evento proporciona cinco regiones de evento auditivo combinado (los límites de evento se “ORed” juntos).

La figura 4 muestra un ejemplo de una señal de entrada de cuatro canales de señal de entrada. Los canales 1 y 4 contienen cada uno tres eventos auditivos y los canales 2 y 3 contienen cada uno dos eventos auditivos. Los límites de evento auditivo combinados para los bloques de datos simultáneos a través de los cuatro canales están ubicados en los números de muestras 512, 1024, 1536, 2560 y 3072 como se indica en la parte inferior de la figura 4.

En principio, el audio procesado puede ser digital o analógico y no necesita dividirse en bloques. Sin embargo, en aplicaciones prácticas, las señales de entrada son posiblemente uno o más canales de audio digital representados por muestras en las que las muestras consecutivas de cada canal se dividen en bloques de, por ejemplo 4096 muestras (como en los ejemplos de las figuras 1, 3 y 4, más arriba). En realizaciones prácticas establecidas en el presente documento, los eventos auditivos se determinan examinando bloques de datos de muestras de audio que preferiblemente representan aproximadamente 20 ms de audio o menos, que se cree que es el evento auditivo más corto reconocible por el oído humano. Así, en la práctica, es probable que los eventos auditivos se determinen examinando bloques de, por ejemplo, 512 muestras que corresponden a unos 11,6 ms de de audio de entrada a una frecuencia de muestreo de 44.1 kHz dentro de grandes bloques de datos de la muestra de audio. Sin embargo, a lo largo de todo este documento se hace referencia para los “bloques” más que para “sub-bloques” cuando se refiere al examen de los segmentos de datos de audio con el propósito de detectar límites de evento auditivo. Debido a que los datos de muestra de audio se examinan en bloques, en la práctica los límites de inicio temporal y punto de parada temporal del evento auditivo necesariamente coincidirán con límites de bloque. Hay una compensación entre los requisitos de procesamiento en tiempo real (ya que los bloques grandes requieren menos gastos de procesamiento) y la resolución de ubicación del evento (los bloques pequeños proporcionan información más detallada de la ubicación de evento auditivo).

Otros aspectos de la invención se apreciarán y entenderán cuando se haya leído y entendido la descripción detallada de la invención.

Breve descripción de los dibujos

La figura 1 es una forma de onda idealizada de un único canal de música orquestal que ilustra auditivo.

La figura 2 es un diagrama esquemático conceptual idealizado que ilustra el concepto de dividir anchos de banda de audio completos en sub-bandas de frecuencia con el fin de identificar eventos auditivos de sub-banda. La escala

horizontal son muestras y la escala vertical es frecuencia.

La figura 3 es una serie de formas de onda idealizadas en dos canales de audio, que muestra eventos de audio en cada canal y eventos de audio combinados a través de los dos canales.

La figura 4 es una serie de forma de ondas idealizadas en cuadro canales de audio que muestra eventos de audio en cada canal y eventos de audio combinados a través de los cuatro canales.

La figura 5 es un diagrama de flujo que muestra la extracción de ubicaciones de evento de audio y la extracción opcional de sub-bandas dominantes de una señal de audio de acuerdo con la presente invención.

La figura 6 es una representación esquemática conceptual que describe un análisis espectral de acuerdo con la presente invención.

Descripción detallada de las realizaciones preferidas

De acuerdo con una realización de un aspecto de la presente invención, el análisis de escenas auditivas está compuesto de tres pasos generales de procesamiento como se muestra en una porción de la figura 5. El primer paso 5-1 (“Realizar el Análisis Espectral”) toma una señal de audio de dominio tiempo, la divide en bloques y calcula un perfil espectral o contenido espectral para cada uno de los bloques. El análisis espectral transforma la señal de audio en el dominio frecuencia a corto plazo. Esto puede ejecutarse usando cualquier banco de filtros, ya sea basado en transformadas o bancos de filtros de paso de banda, y en el espacio de frecuencia bien lineal o bien combado (tal como la escala Bark o banda crítica, que mejor se aproxima a las características del oído humano). Con cualquier banco de filtros existe una compensación entre tiempo y frecuencia. Mayor tiempo de resolución y por tanto intervalos de tiempo más cortos, conduce a resolución de frecuencia más baja. Mayor resolución de frecuencia, y por tanto sub-bandas más estrechas, conducen a intervalos de tiempo más largos.

El primer paso, ilustrado conceptualmente en la figura 6, calcula el contenido espectral de segmentos sucesivos de tiempo de la señal de audio. En una realización práctica. El tamaño del bloque ASA es 512 muestras de la señal de audio de entrada. En el segundo paso 5-2, las diferencias en el contenido espectral de bloque a bloque están determinadas (“Realizar mediciones de diferencia de perfil espectral”). Así, el segundo paso calcula la diferencia en el contenido espectral entre segmentos de tiempo sucesivos de la señal de audio. Como se discutía anteriormente, un indicador poderoso del comienzo o final de un evento auditivo percibido se cree que es un cambio en el contenido espectral. En el tercer paso 5-3 (“Identificar la ubicación de los límites del evento auditivo”), cuando la diferencia espectral entre un bloque de perfil espectral y el siguiente es mayor que un umbral, el límite del bloque se toma como un límite de evento auditivo. El segmento de audio entre límites consecutivos constituye un evento auditivo. Así, el tercer paso establece un límite de evento auditivo entre sucesivos segmentos de tiempo cuando la diferencia en el contenido del perfil espectral entre tales segmentos de tiempo sucesivos supera un umbral, definiendo así eventos auditivos. En esta realización, los límites de evento auditivo definen eventos auditivos con una longitud que es un múltiplo integral de bloques de perfil espectral con una longitud mínima de un bloque de perfil espectral (512 muestras en este ejemplo). En principio, los límites del evento no necesitan estar tan limitados. Como alternativa a las realizaciones prácticas aquí discutidas, el tamaño de bloque de entrada puede variar, por ejemplo, con el fin de ser esencialmente del tamaño de un evento auditivo.

Las ubicaciones de los límites del evento se pueden almacenar como una caracterización de información reducida o “firma” y con el formato deseado, como se muestra en el paso 5-4. Un paso de proceso opcional 5-5 (“identificar sub-banda dominante”) usa el análisis espectral del paso 5-1 para identificar una sub-banda de frecuencia dominante que también se puede almacenar como parte de la firma. La información de sub-banda dominante se puede combinar con la información de límite de evento para definir un rasgo de cada evento auditivo.

Segmentos bien de superposición o bien sin superposición del audio se pueden disponer en ventanas y utilizar para computar perfiles espectrales del audio de entrada. La superposición da como resultado una resolución más fina, en lo relativo a la ubicación de elementos auditivos, y también hace que sea menos probable que se pierda un evento, tal como un transitorio. Sin embargo, la superposición también aumenta la complejidad computacional. Así, la superposición se puede omitir. La figura 6 muestra una representación conceptual de 512 bloques de muestra sin superposición que se están disponiendo en ventanas y transformando en el dominio frecuencia mediante la transformada de Fourier discreta (DFT del inglés *Discrete Fourier Transform*). Cada bloque se puede disponer en ventanas y transformar en el dominio frecuencia, por ejemplo utilizando la DFT, implementada preferiblemente como una transformada de Fourier rápida (FFT del inglés *Fast Fourier Transform*) para la velocidad.

Las siguientes variables pueden usarse para computar el perfil espectral del bloque de entrada.

N = número de muestras de la señal de entrada

M = número de muestras dispuestas en ventanas en un bloque, usadas para calcular el perfil espectral

P = número de muestras de solapamiento de cálculo espectral

Q = número de ventanas/regiones computadas.

En general, se puede utilizar cualquier número entero para las variables anteriores. Sin embargo, la implementación será más eficiente si M se fija igual a una potencia de 2 de modo que se pueden utilizar FFT normalizadas para los cálculos del perfil espectral. Además, si N, M y P se eligen de tal manera que Q sea un número entero, esto evitará un audio infradimensionado o sobredimensionado al final de las N muestras. En una realización práctica del proceso de análisis de escenas auditivas, los parámetros enumerados se pueden establecer:

M = 512 muestras (o bien 11.6 ms a 44.1 kHz)

P = 0 muestras (sin superposición)

Los valores arriba enumerados se determinaron experimentalmente y se encontró generalmente que identificaban con suficiente exactitud la ubicación y duración de eventos auditivos. Sin embargo, estableciendo el valor de P a 256 muestras (50% superposición) en lugar de cero muestras (sin superposición) se ha encontrado que es útil para identificar algunos eventos difíciles de encontrar. Aunque se pueden usar muchos tipos de ventanas diferentes para minimizar artefactos espectrales debido a la disposición en ventanas, la ventana usada en los cálculos de perfil espectral es una ventana Hanning de M puntos, Kaiser-Bessel u otra apropiada, preferiblemente no rectangular. Los valores arriba indicados y una ventana tipo Hanning se seleccionaron tras un análisis experimental extenso y han demostrado proporcionar excelentes resultados a través de un amplio rango de material de audio. La ventana no rectangular se prefiere para el procesamiento de señales de audio con un contenido predominante de baja frecuencia. La ventana rectangular produce artefactos espectrales que pueden causar una detección incorrecta de eventos. A diferencia de ciertas aplicaciones codificador/decodificador (codec) donde un proceso global de superposición/adición debe proporcionar un nivel constante, tal restricción no se aplica aquí y la ventana se puede elegir para características como la resolución de frecuencia / tiempo y el rechazo de parada de banda.

En el paso 5-1 (figura 5) el espectro de cada bloque de M muestras se puede computar disponiendo en ventanas los datos por una ventana Hanning de M puntos, Kaiser-Bessel u otra apropiada, convirtiendo al dominio frecuencia usando una transformada de Fourier rápida de M puntos y calculando la magnitud de los coeficientes FFT complejos. Los datos resultantes se normalizan de modo que la mayor magnitud se ajusta a la unidad, y la formación normalizada de M números se convierte en el dominio logarítmico. La formación no necesita convertirse en el dominio logarítmico, pero la conversión simplifica los cálculos de la diferente medida en el paso 5-2. Además, el dominio logarítmico es más compatible con la naturaleza del sistema auditivo humano. Los valores del dominio logarítmico resultante tienen un alcance de menos infinito a cero. En una realización práctica, se puede imponer un límite inferior en el intervalo de valores; el límite se puede fijar, por ejemplo -60 dB, o ser dependiente de la frecuencia para reflejar la baja audibilidad de sonidos tranquilos en frecuencias bajas y muy altas. (Hay que tener en cuenta que sería posible reducir el tamaño de la formación a M/2 en la que FFT representa frecuencias negativas al igual que positivas).

El paso 5-2 calcula una medida de la diferencia entre el espectro de bloques adyacentes. Para cada bloque, cada uno de los coeficientes espectrales M (logaritmo) del paso 5-1 se resta del correspondiente coeficiente del bloque anterior, y la magnitud de la diferencia se calcula (el signo se ignora). Estas M diferencias se suman entonces a un número. Por lo tanto, para un segmento de tiempo continuo de audio, que contiene Q bloques, el resultado es una formación de Q números positivos, uno para cada bloque. Cuanto mayor sea el número, más difiere un bloque en espectro del bloque anterior. Esta medida de diferencia también se puede expresar como una diferencia promedio por coeficiente espectral dividiendo la medida de diferencia entre el número de coeficientes espectrales usados en la suma (en este caso coeficientes M).

El paso 5-3 identifica las ubicaciones de límites de eventos auditivos aplicando un umbral a la formación de medidas de diferencia del paso 5-2 con un valor umbral. Cuando una medida de diferencia supera un umbral, el cambio en el espectro se considera suficiente para indicar un nuevo evento y el número de bloque del cambio se registra como un límite de evento. Para los valores de M y P dados anteriormente y para los valores de dominio de logaritmo (en el paso 5-1) expresado en unidades de dB, el umbral se puede establecer igual a 2500 si se compara todo el conjunto de la magnitud FFT (incluyendo la parte reflejada) o 1250 si se compara la mitad de FFT (como se señaló anteriormente, el FFT representa frecuencias tanto negativas como positivas – para la magnitud del FFT, una es la imagen especular de la otra). Este valor fue elegido experimentalmente y proporciona una buena detección del límite de evento auditivo. Este valor de parámetro puede cambiarse para reducir (aumentar el umbral) o aumentar (disminuir el umbral) la detección de eventos.

Para una señal de audio que consista en Q bloques (de un tamaño de M muestras), la salida del paso 5-3 de la figura 5 se puede almacenar y formatear en el paso 5-4 como una formación *B(q)* de información que representa la ubicación de los límites de evento auditivo donde *q* = 0, 1,… Q-1. Para un tamaño de bloque de M = 512 muestras, superposición de P = 0 muestras y una velocidad de señal de muestreo de 44.1kHz, la función de análisis de escenas auditivas 2 imprime aproximadamente 86 valores por segundo. La formación *B(q)* puede almacenarse como

una firma, de tal manera que, en su forma básica, sin la información opcional de frecuencia de sub-banda dominante del paso 5-5, la firma de la señal de audio es una formación *B(q)* que representa una cadena de límites de evento auditivo.

Identificar sub-banda dominante (opcional)

Para cada bloque, un paso adicional opcional en el procesamiento de la figura 5 es extraer información de la señal de audio que denota la frecuencia dominante "sub-banda" del bloque (conversión de los datos en cada bloque a los resultados del dominio de la frecuencia en información dividida en sub-bandas de frecuencia). Esta información basada en el bloque se puede convertir en información basada en evento auditivo, de forma que la sub-banda de frecuencia dominante se identifique para cada evento auditivo. Tal información para cada evento auditivo proporciona información con respecto al evento auditivo en sí y puede ser útil para proporcionar una representación de información reducida de la señal de audio más detallada y única. El empleo de información de sub-banda dominante es más apropiado en el caso de determinar eventos auditivos de ancho de banda completo en lugar de casos en los cuales el audio se rompe en sub-bandas y los eventos auditivos están determinados por cada sub-banda.

La sub-banda dominante (mayor amplitud) se puede elegir de entre una pluralidad de sub-bandas, tres o cuatro, por ejemplo, que están dentro del alcance o banda de frecuencias donde el oído humano es más sensible. Alternativamente, se pueden utilizar otros criterios para seleccionar las sub-bandas. El espectro se puede dividir, por ejemplo, en tres sub-bandas. Útiles alcances de frecuencia para las sub-bandas son (estas frecuencias en particular no son críticas):

Sub-banda 1	300 Hz hasta 550 Hz
Sub-banda 2	550 Hz hasta 2000 Hz
Sub-banda 3	2000 Hz hasta 10.000 Hz

Para determinar la sub-banda dominante, el cuadrado de la magnitud del espectro (o el espectro de la magnitud del poder) se suma para cada sub-banda. Esta suma resultante para cada sub-banda se calcula y se elige la mayor. Las sub-bandas también pueden ponderarse antes de seleccionar la mayor. La ponderación puede adoptar la forma de dividir la suma para cada sub-banda por el número de valores espectrales en la sub-banda, o alternativamente puede adoptar la forma de una adición o multiplicación para enfatizar la importancia de una banda sobre otra. Esto puede resultar útil donde algunas sub-bandas tienen más energía que el promedio de otras sub-bandas pero son menos importantes perceptualmente.

Teniendo en cuenta una señal de audio que consiste en Q bloques, la salida del procesamiento de la sub-banda dominante es una formación *DC(q)* de información que representa la sub-banda dominante de cada bloque *(q = 0, 1,… Q-1).* Preferiblemente, la formación *DS(q)* está formateada y almacenada en la firma junto con la formación *B(q).* Así, con la información opcional de sub-banda dominante, la firma de la señal de audio consiste en dos formaciones *B(q)* y *DS(q),* que representan, respectivamente, una cadena de límites de evento auditivo y una sub-banda de frecuencia dominante dentro de cada bloque, de la cual se puede determinar si se desea la sub-banda de frecuencia dominante para cada evento auditivo. Así, en un ejemplo idealizado, las dos formaciones podrían tener los siguientes valores (para un caso en el que haya tres posibles sub-bandas dominantes).

1 0 1 0 0 0 1 0 0 1 0 0 0 0 0 1 0 (Límites de Evento)

1 1 2 2 2 2 1 1 1 3 3 3 3 3 3 1 1 (Sub-bandas Dominantes)

En la mayoría de los casos, la sub-banda dominante permanece igual dentro de cada evento auditivo, como se muestra en este ejemplo, o tiene un valor promedio si no es uniforme para todos los bloques dentro del evento. Así, una sub-banda dominante se puede determinar para cada evento auditivo y la formación *DS(q)* se puede modificar para proporcionar que la misma sub-banda dominante se asigne a cada bloque dentro de un evento.

El proceso de la figura 5 se puede representar más generalmente por las disposiciones equivalentes de las figuras 7, 8 y 9. En la figura 7, se aplica una señal de audio en paralelo a una función "Identificar Eventos Auditivos" o paso 7-1 que divide la señal de audio en eventos auditivos, cada uno de los cuales tiende a ser percibido como separado y distinto, y a una función opcional "Identificar Características de Eventos Auditivos" o paso 7-2. El proceso de la figura 5 se puede emplear para dividir la señal de audio en eventos auditivos o se puede emplear otro proceso apropiado. La información de evento auditivo que puede ser una identificación de los límites de evento auditivo, determinada por la función del paso 7-1 se almacena y formatea, como se desee, por una función "Almacenar y Formatear" o paso 7-3. La función opcional "Identificar Características" o paso 7-3 también recibe la información del evento auditivo. La función "Identificar Características" o paso 7-3 puede caracterizar algunos o todos los eventos auditivos por una o más características. Tales características pueden incluir una identificación de la sub-banda dominante del evento auditivo, como se describe en relación con el proceso de la figura 5. Las características también pueden

incluir uno o más descriptores de audio MPEG-7, incluyendo, por ejemplo, una medida de potencia del evento auditivo, una medida de amplitud del evento auditivo, una medida de planicidad espectral del evento auditivo, y si el evento auditivo es sustancialmente silencioso. Las características también pueden incluir otras características tales como si el evento auditivo incluye un transitorio. Características para uno o más eventos auditivos también se reciben por la función “Almacenar y Formatear” o paso 7-3 y se almacena y formatea junto con la información de evento auditivo.

Alternativas a la disposición de la figura 7 se muestra en las figuras 8 y 9. En la figura 8, la señal de entrada de audio no se aplica directamente a la función “Identificar Características” o paso 8-3, pero recibe información de la función “Identificar Eventos Auditivos” o paso 8-1. La disposición de la figura 5 es un ejemplo específico de tal disposición. En la figura 9, las funciones o pasos 9-1, 9-2 y 9-3 están dispuestas en series.

Los detalles de esta realización práctica no son críticos. Otros modos de calcular el contenido espectral de sucesivos segmentos de tiempo de la señal de audio, calculan las diferencias entre sucesivos segmentos de tiempo, y establecen los límites de evento auditivo en los límites respectivos entre segmentos de tiempo cuando la diferencia en el contenido de perfil espectral entre tales segmentos de tiempo sucesivos supera un umbral que se puede emplear.

La presente invención y sus varios aspectos se pueden implementar como funciones de software realizadas en procesadores de señales digitales, ordenadores digitales programados de uso general, y/u ordenadores digitales para un propósito especial. Interfaces entre corrientes de señales analógicas y digitales se pueden realizar en hardware apropiado como funciones en software y/o firmware.

REIVINDICACIONES

1. Un método para dividir cada uno de los múltiples canales de señales de audio digital en eventos auditivos, que comprende:

detectar cambios en el contenido espectral con respecto al tiempo en la señal de audio en cada uno de los canales (5.2), donde los cambios en el contenido espectral se calculan en el dominio logaritmo,

identificar los límites de evento auditivo en la señal de audio de un canal (5-3), donde cada límite es la respuesta a un cambio en el contenido espectral con respecto al tiempo en el canal que supera un umbral de tal manera que se obtiene un conjunto de límites de evento auditivo para cada canal, y cada segmento de audio en un canal entre límites consecutivos constituye un evento auditivo, e

identificar un límite de evento auditivo combinado para los canales en respuesta a la identificación de un límite de evento auditivo en cualquier canal.

2. Un método de acuerdo con la reivindicación 1, en el que el audio en respectivos canales representa respectivas direcciones en el espacio.

3. Un método de acuerdo con la reivindicación 1, en el que el audio en respectivos canales representa bandas de frecuencia de una señal de audio.

4. Un método de acuerdo con una cualquiera de las reivindicaciones 1-3, en el que dichos cambios de detección en el contenido espectral con respecto al tiempo en la señal de audio en cada uno de los canales incluye dividir la señal de audio en bloques de tiempo y convertir los datos en cada bloque al dominio frecuencia.

5. Un método de acuerdo con la reivindicación 4, en el que dichos cambios de detección en el contenido espectral con respecto al tiempo en la señal de audio en cada uno de los canales detectan los cambios en el contenido espectral entre sucesivos bloques de tiempo de la señal de audio en cada uno de los canales.

6. El método de la reivindicación 5, en el que los datos de audio en bloques de tiempo consecutivos se representan por coeficientes y dichos cambios de detección en el contenido espectral entre sucesivos bloques de tiempo de la señal de audio en cada uno de los canales incluye:

convertir dichos coeficientes al dominio logaritmo, y

substraer coeficientes de un bloque de los coeficientes correspondientes de un bloque adyacente.

7. El método de la reivindicación 6, en el que detectar cambios en el contenido espectral entre sucesivos bloques de tiempo de la señal de audio en cada uno de los canales incluye además:

sumar las magnitudes de las diferencias resultantes de sustraer coeficientes de un bloque de coeficientes correspondientes de un bloque adyacente, y

comparar las magnitudes sumadas con un umbral.

8. El método de reivindicación 7, en el que un límite de evento auditivo se identifica cuando las magnitudes sumadas superan dicho umbral.

9. El método de la reivindicación 4, en el que dicho método comprende además asignar una característica a uno o más de los eventos auditivos.

10. El método de la reivindicación 5, en el que características asignables a uno o más de los eventos auditivos incluyen uno o más de: la sub-banda dominante del espectro de frecuencia del evento auditivo, una medida de potencia del evento auditivo, una medida de amplitud del evento auditivo, una medida de planicidad espectral del evento auditivo, dónde el evento auditivo es sustancialmente silencioso, y dónde el evento auditivo incluye un transitorio.

11. El método de la reivindicación 10, que comprende además formatear y almacenar los límites de evento auditivo e identificar las características asignadas a eventos auditivos.

12. El método de cualquiera de las reivindicaciones 1 a 4, que comprende además formatear y almacenar los límites de evento auditivo.

13. El método de la reivindicación 4, en el que dichos cambios de detección en el contenido espectral con respecto al tiempo en la señal de audio en cada uno de los canales detecta además cambios en amplitud entre sucesivos

bloques de tiempo de la señal de audio en cada uno de los canales.

14. El método de la reivindicación 3, en el que dichos cambios en la amplitud se detectan por el grado en que se normalizan los datos del dominio frecuencia.

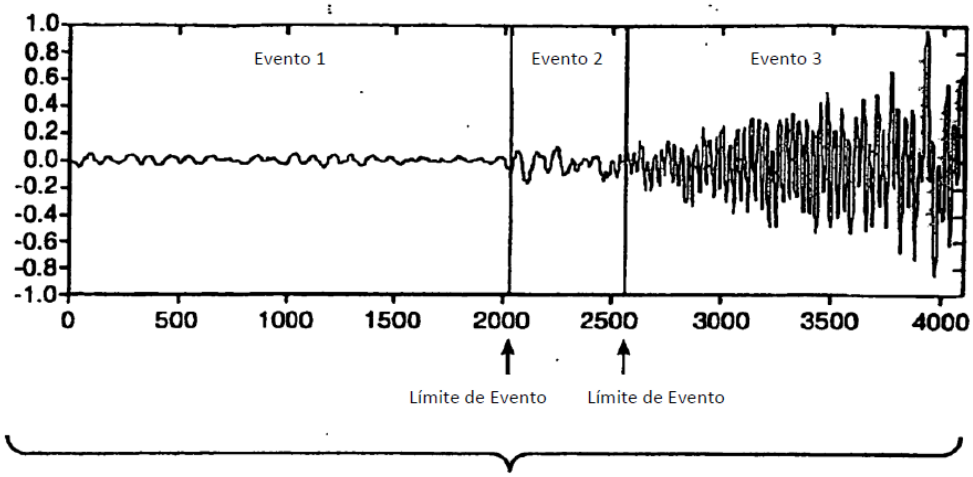
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Evento 1
Evento 2
Evento 3
0
500
1000
1500
2000
2500
3000
3500
4000
Límite de Evento
Límite de Evento

FIG._1

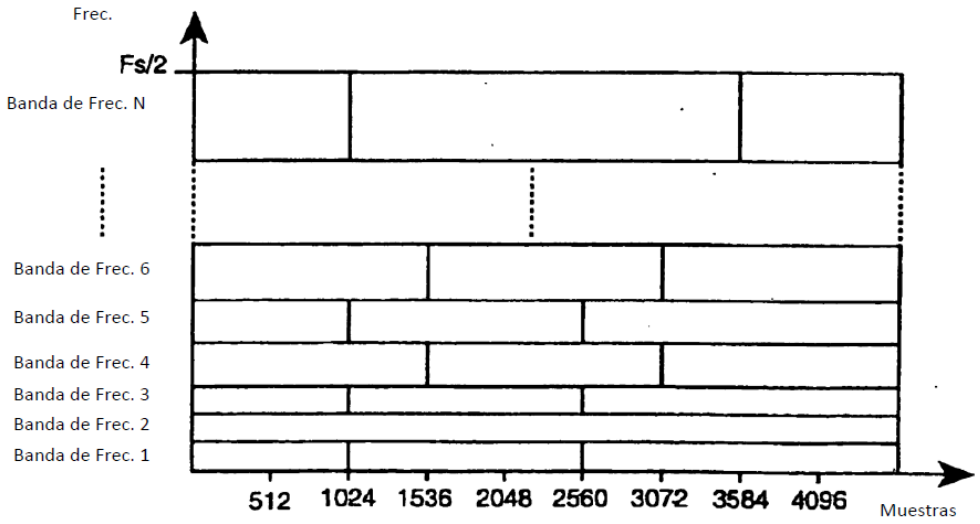
Frec.
Fs/2
Banda de Frec. N
Banda de Frec. 6
Banda de Frec. 5
Banda de Frec. 4
Banda de Frec. 3
Banda de Frec. 2
Banda de Frec. 1
512
1024
1536
2048
2560
3072
3584
4096
Muestras

FIG._2

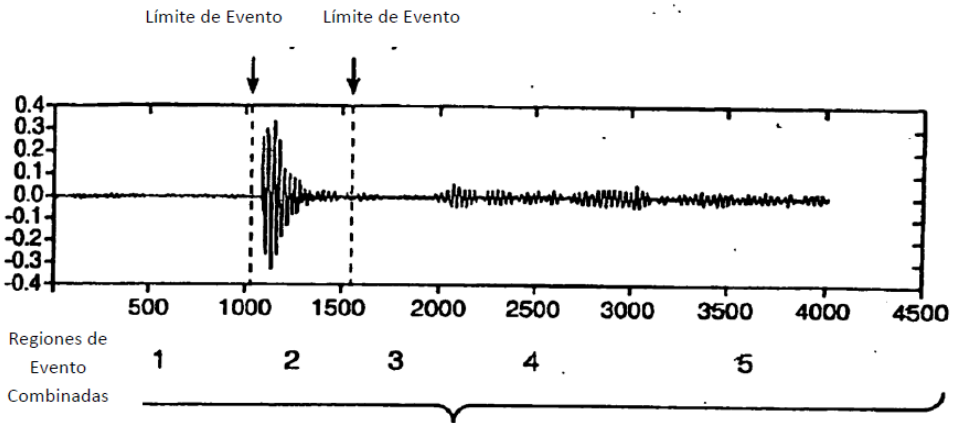
Límite de Evento
Límite de Evento
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
500
1000
1500
2000
2500
3000
3500
4000
4500
Regiones de Evento Combinadas
1
2
3
4
5

FIG._3A

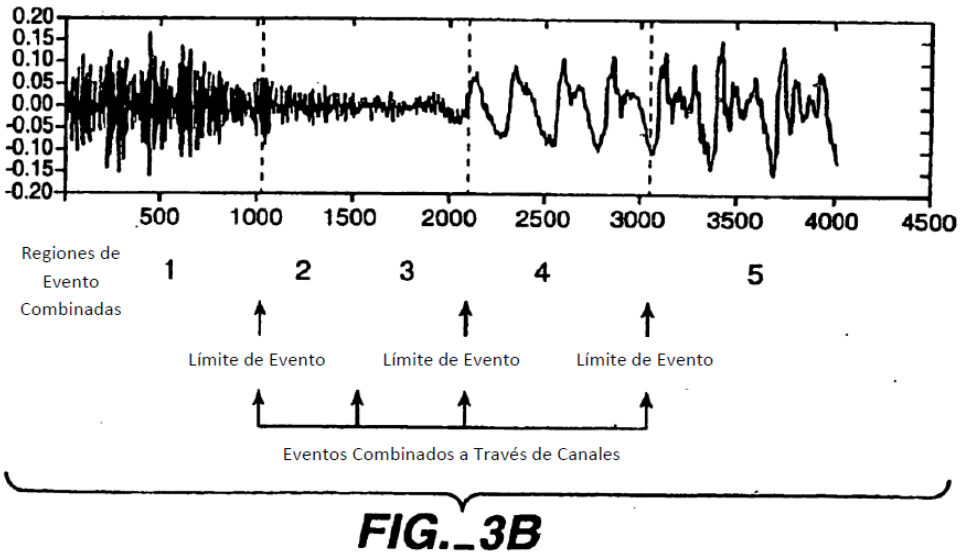
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
500
1000
1500
2000
2500
3000
3500
4000
4500
Regiones de Evento Combinadas
1
2
3
4
5
Límite de Evento
Límite de Evento
Límite de Evento
Eventos Combinados a Través de Canales

FIG._3B

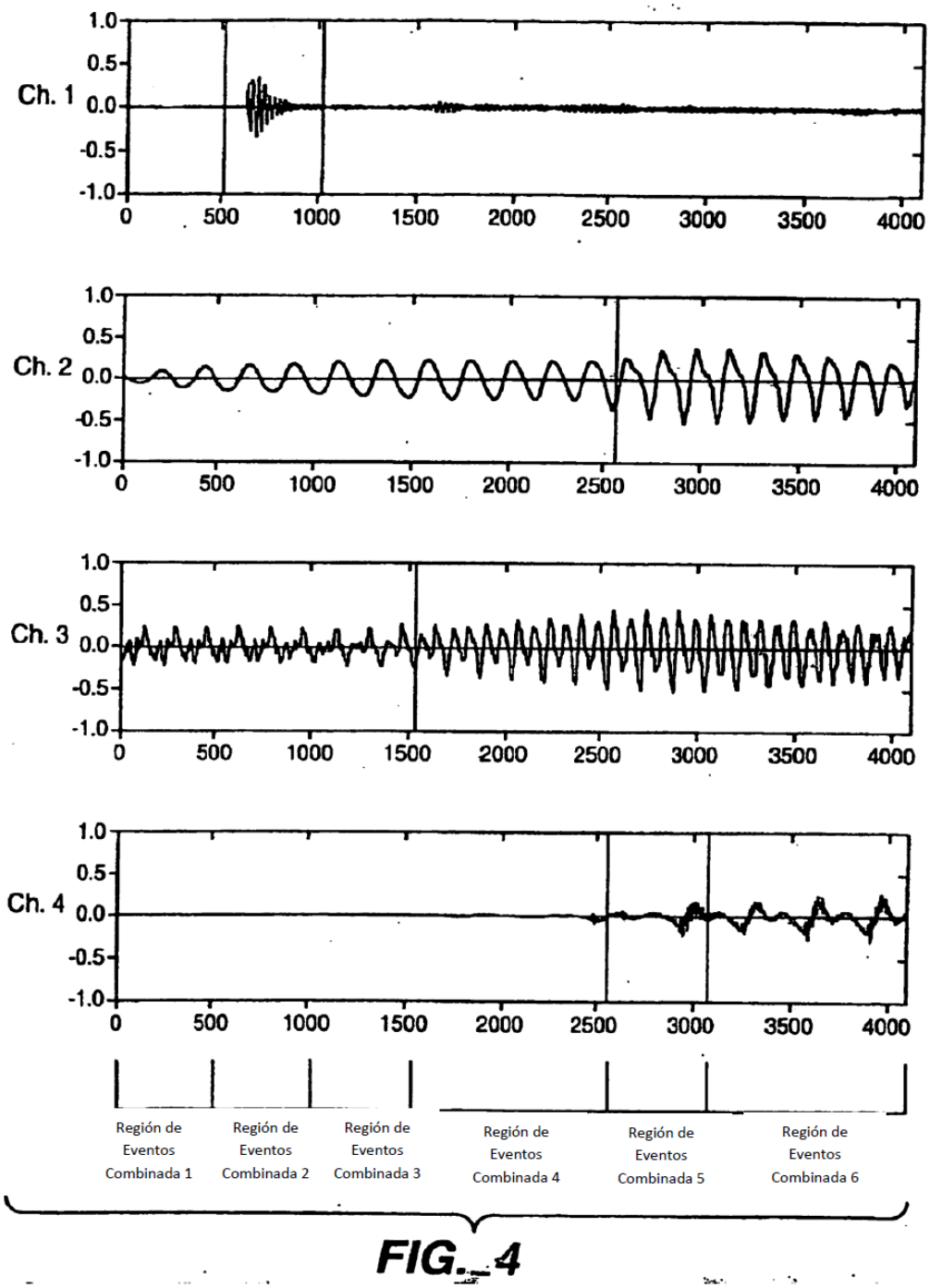
Ch. 1
Ch. 2
Ch. 3
Ch. 4
1.0
0.5
0.0
-0.5
-1.0
0
500
1000
1500
2000
2500
3000
3500
4000
Región de Eventos Combinada 1
Región de Eventos Combinada 2
Región de Eventos Combinada 3
Región de Eventos Combinada 4
Región de Eventos Combinada 5
Región de Eventos Combinada 6

FIG._4

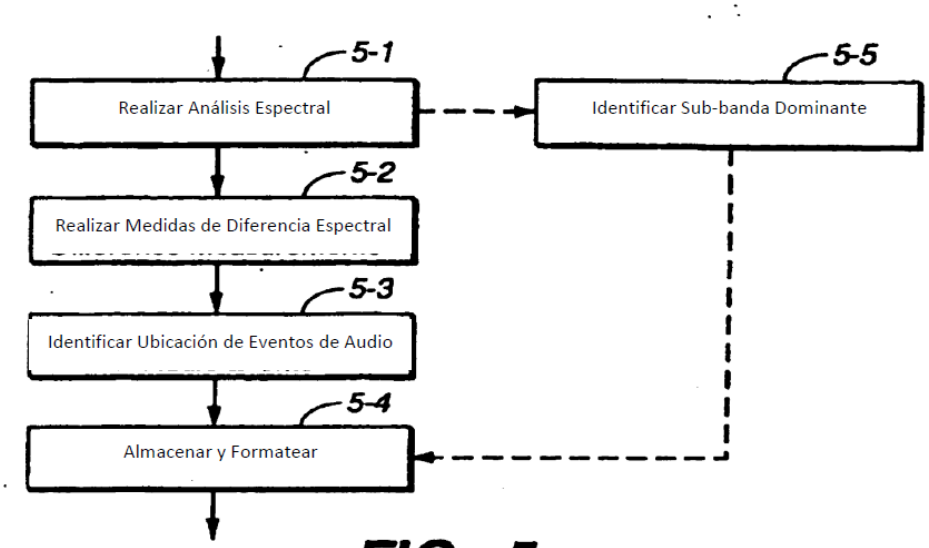
5-1
Realizar Análisis Espectral
5-5
Identificar Sub-banda Dominante
5-2
Realizar Medidas de Diferencia Espectral
5-3
Identificar Ubicación de Eventos de Audio
5-4
Almacenar y Formatear

FIG._5

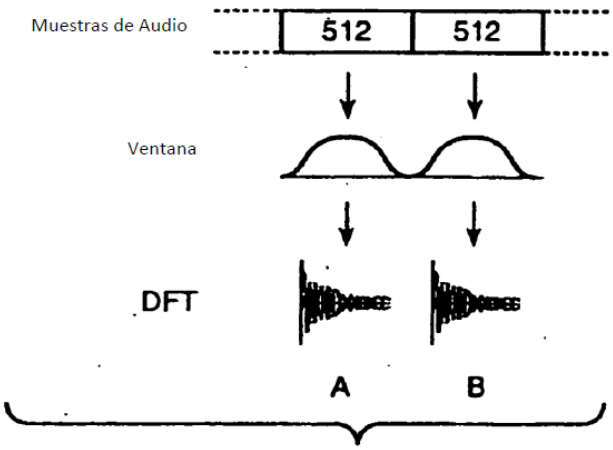
Muestras de Audio
512
512
Ventana
DFT
A
B

FIG._6

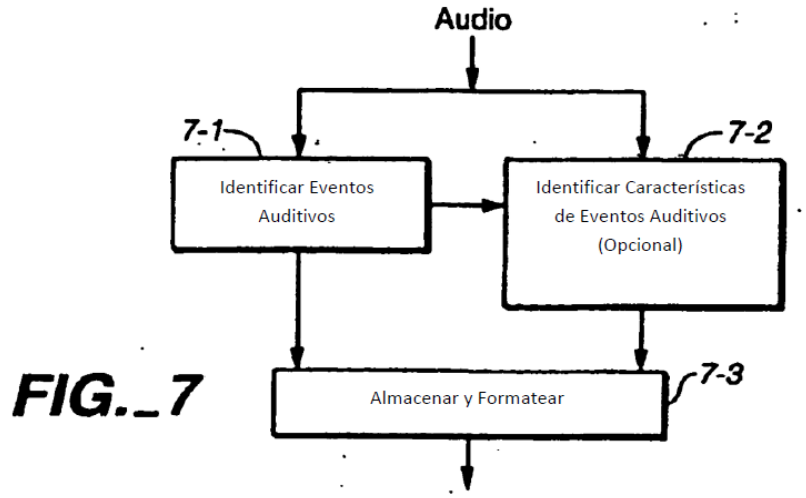

FIG._7

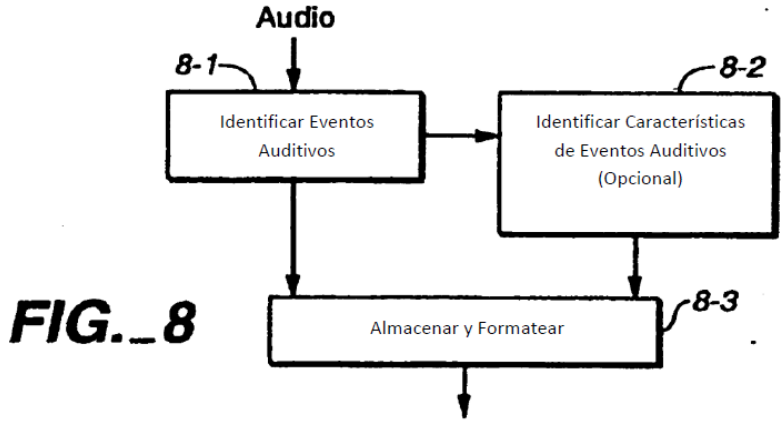

FIG._8

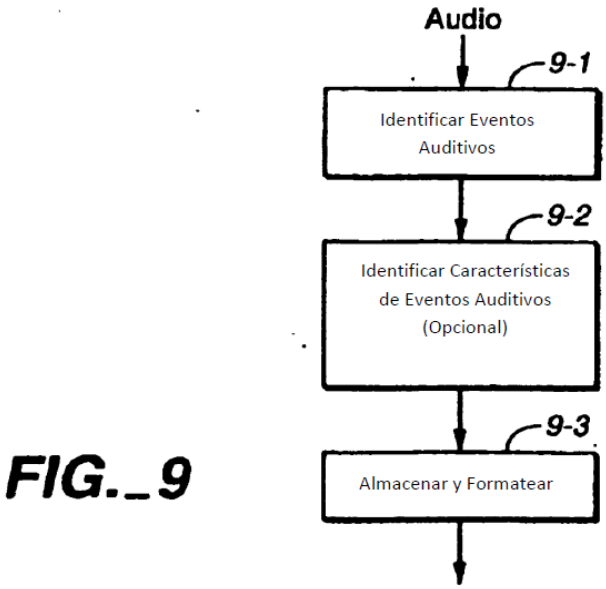

FIG._9