(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 986 239**

(21) Número de solicitud: 202430765

(51) Int. CI.:

A61F 11/14 (2006.01)
G10K 11/16 (2006.01)
H04R 1/10 (2006.01)

(12) SOLICITUD DE PATENTE A1

(22) Fecha de presentación:

26.09.2024

(43) Fecha de publicación de la solicitud:

08.11.2024

(71) Solicitantes:

UNIVERSITAT POLITÈCNICA DE VALÈNCIA (100.0%)
C/ ALMIRALL CADARSO, 26, BAJO
46005 VALENCIA (Valencia) ES

(72) Inventor/es:

LÓPEZ MONFORT, José Javier;
BANCHERO MARTÍNEZ, Lucas y
ALVAREZ MARTÍNEZ, Ariel

(74) Agente/Representante:

SÁNCHEZ MARGARETO, Carolina

(54) Título: **Dispositivo y procedimiento de protección auditiva**

(57) Resumen:

Dispositivo de protección auditiva, formado por unos cascos (1), unos micrófonos (2) extensores y unos transductores (3) interiores, gestionados por una unidad de control (4), caracterizado por que los cascos (1) son de protección auditiva, configurados para eliminar el sonido ambiental, y la unidad de control (4) está configurada para seleccionar del sonido captado por los micrófonos (2) las señales relevantes y emitirlas por los transductores (3). La unidad de control (4) puede estar configurada para adaptar la señal emitida al usuario para simular la proveniencia del sonido relevante.

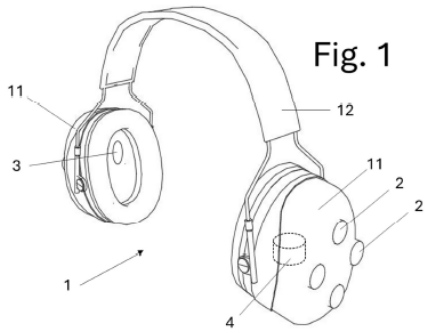


Fig. 1

ES 2 986 239 A1

DESCRIPCIÓN

Dispositivo y procedimiento de protección auditiva

SECTOR DE LA TÉCNICA

La presente invención se refiere a un dispositivo de protección auditiva para ambientes ruidosos que permite sin embargo la transmisión al usuario de los sonidos de interés del entorno, incluida la voz de otros compañeros. También se refiere al procedimiento utilizado.

Es aplicable en el campo de la protección personal en industria y otros ámbitos.

ESTADO DE LA TÉCNICA

Los cascos de protección auditiva convencionales permiten suprimir por completo los sonidos externos, logrando un pleno aislamiento acústico pasivo. La protección auditiva pasiva tiene una alta capacidad para atenuar los niveles de ruido percibidos por el usuario, ofreciendo una barrera física entre el oído y el entorno ruidoso.

Los cascos de protección auditiva se clasifican comúnmente en función de su capacidad de atenuación del ruido en diferentes bandas de frecuencia, identificadas como HML. Estas bandas se definen como: H (High) para frecuencias altas (más de 2000 Hz), M (Medium) para frecuencias medias (entre 500 y 2000 Hz) y L (Low) para frecuencias bajas (por debajo de 500 Hz). Los niveles de atenuación en estas bandas se expresan en decibelios (dB).

Se considera que un casco ofrece un aislamiento alto cuando la atenuación en la banda de frecuencias bajas supera los 27 dB, dado que las frecuencias bajas suelen ser las más difíciles de atenuar de manera efectiva. Un ejemplo representativo de un casco con aislamiento elevado comercial de una marca reconocida suele presentar unos valores de referencia de H=37 dB, M=35 dB y L=27 dB.

Sin embargo, si ese aislamiento es excesivo podría omitir señales auditivas cruciales para la seguridad del trabajador. Por eso, la mayor parte de las normativas limitan el aislamiento que los cascos pueden ofrecer (por ejemplo, véase la norma DIN EN 352).

Se conocen sistemas para permitir la transmisión de sonidos relevantes al usuario. Así, en US11890168B2 se describe un sistema para la protección auditiva y la mejora de la conciencia situacional en ambientes ruidosos. No utiliza cascos de protección auditiva con aislamiento pleno, sino que se basa en el procesamiento de señal y la inteligencia artificial, creando una protección auditiva activa. Así, funciona eliminando ruidos mediante el procesamiento de señales acústicas en tiempo real. Este método, aunque eficaz, no proporciona un aislamiento completo debido a que las técnicas de cancelación activa de ruido no son especialmente efectivas a altas frecuencias. Más aún, un fallo repentino del sistema produce un choque auditivo en el usuario.

El solicitante no conoce ningún dispositivo que pueda ser considerado similar a la invención.

BREVE EXPLICACIÓN DE LA INVENCIÓN

La invención consiste en un dispositivo de protección auditiva según las reivindicaciones independientes y cuyas variantes resuelven los problemas del estado de la técnica. La invención también consiste en el procedimiento utilizado.

Es un sistema integral de hardware y software diseñado para mejorar la seguridad y el confort acústico de trabajadores expuestos a sonidos de alta intensidad.

El sistema propuesto aborda esta problemática mediante un sistema combinado mediante unos cascos de protección auditiva pasiva y la integración de micrófonos externos y transductores internos. Los micrófonos capturan el entorno sonoro, ignorando los sonidos no deseados y destacando aquellos relevantes, como avisos acústicos, voces de otras personas que se dirigen al usuario, así como cualquier sonido relevante en la seguridad. A través de un software y un hardware específicamente desarrollado para esta tarea, los sonidos de interés comentados son detectados, aislados del resto del ruido y localizados en el espacio alrededor del usuario. Por ejemplo, pueden presentar una señal de alarma o el habla de un compañero de trabajo, eliminando cualquier ruido no deseado gracias al proceso de extracción de la señal de interés frente a otros ruidos.

A continuación, son reproducidos mediante los transductores integrados en el interior de los cascos de protección auditiva, proporcionando la sensación de que provienen de la misma dirección donde se han producido. El casco o cascos de protección auditiva

elimina prácticamente todo el sonido exterior para que no penetre en los oídos del usuario, por lo que solo escucha lo reproducido por los transductores internos. Este enfoque proporciona una protección auditiva muy alta, que se loga de manera pasiva por el aislamiento físico del casco y que puede ser todo lo grande que se desee. Los sonidos relevantes son aislados y reproducidos por el sistema descrito, siendo mejor escuchados cuanta mayor protección ofrezcan los cascos. La seguridad del trabajador es, por tanto, mucho más elevada que en sistemas únicamente pasivos o de cancelación activa de ruido. En conjunto, esta solución representa un enfoque avanzado y completo para la protección auditiva en entornos laborales con altos niveles de ruido, marcando un hito significativo en la mejora de la seguridad y el confort del trabajador en ambientes laborales desafiantes.

El sonido capturado por los micrófonos contiene una gran cantidad de señales y ruidos ambientales. Por ello, utilizando la información previa proporcionada por un modelo de transformadores ("*Transformers*" en el lenguaje habitual de la técnica), que actúa como un disparador y clasificador de sonidos, se puede identificar y categorizar los diferentes tipos de sonidos presentes en la señal y seleccionar los relevantes. Esta información es crucial para el proceso de extracción de señales de interés de manera limpia y precisa.

Las pistas de audio limpias y separadas se envían al procesamiento de la siguiente etapa del sistema, que se encargará de la localización y emulación del sonido en los cascos de protección auditiva. En resumen, la extracción de sonido es un componente crucial del sistema que garantiza que el usuario recibe únicamente las señales de interés de manera limpia y sin ruido, mejorando la eficacia y la utilidad del sistema en entornos laborales ruidosos.

La tarea de la extracción de sonido consiste en separar las señales de interés, como alarmas, discursos u otros sonidos específicos que el operador necesita escuchar en tiempo real, eliminando cualquier tipo de ruido no deseado. Para lograr esto, se emplean algoritmos de inteligencia artificial que utilizan la información sobre el tipo de sonido presente en la señal para extraer de manera separada las diferentes pistas de audio correspondientes a los sonidos de interés.

En concreto, el dispositivo de protección auditiva está formado por unos cascos, unos micrófonos exteriores y unos transductores interiores a los cascos, todos ellos gestionados por una unidad de control. Los cascos son de protección auditiva, bloqueando y eliminando en lo posible todo el sonido ambiental. La unidad de control

está configurada para seleccionar del sonido captado por los micrófonos las señales relevantes y emitirlas por los transductores.

Idealmente, la unidad de control está configurada para adaptar la señal emitida al usuario para simular la proveniencia del sonido relevante.

Este procedimiento aplicado en el dispositivo comprende las etapas de:

- Bloquear los sonidos ambientales mediante los cascos.
- Captar los sonidos ambientales mediante unos micrófonos.
- Extraer unos sonidos relevantes de los sonidos ambientales.
- Emitir los sonidos ambientales en unos transductores interiores a los cascos, incluyendo información espacial para su localización, usando diferentes transductores para simular el origen.

Como se ha citado, es deseable incluir una etapa de adaptación de la emisión de los transductores a la fisiología del usuario.

Una ventaja adicional de las realizaciones particulares es el uso de HRTF (funciones de transferencia relacionadas con la cabeza) mediante imágenes de las orejas del usuario y otras partes (cuello, cabeza) tomadas antes de ponerse los transductores. Esta personalización es fundamental porque las HRTF permiten traducir la información de localización de la red neuronal en estímulos que el cerebro interpreta como direccionalidad de sonido. Una HRTF personalizada mejora considerablemente la percepción de la procedencia del sonido, reduciendo errores en la localización subjetiva por parte del trabajador. Esto no solo aumenta la eficacia de la protección auditiva, sino que también mejora la seguridad y la capacidad de respuesta del trabajador en situaciones críticas.

Esta HRTF permite representar el sonido de manera realista y espacialmente precisa en los transductores, en la misma dirección en la que se ha detectado la fuente sonora. Esto significa que el usuario percibirá el sonido como si viniera directamente de la ubicación de la fuente, mejorando significativamente la capacidad de localización y la inmersión auditiva del sistema.

El sistema permite atenuaciones pasivas mejorando las recomendaciones, lo cual redunda en una menor exposición al ruido del trabajador. Las recomendaciones limitan la atenuación de los protectores auditivos para asegurar que el trabajador pueda oír los

sonidos de relevancia. Sin embargo, al detectar, localizar y reproducir esos sonidos mediante la presente invención, esta limitación se puede sortear, mejorando la salud auditiva y el confort del trabajador.

Otras variantes se aprecian en el resto de la memoria.

DESCRIPCIÓN DE LAS FIGURAS

Para una mejor comprensión de la invención, se incluye un apartado de dibujos, que muestra formas de realización ejemplares.

Figura 1: Vista general en perspectiva de un ejemplo de dispositivo.

Figura 2: Vista superior de un segundo ejemplo.

Figura 3: Vista en perspectiva del segundo ejemplo.

Figura 4: Etapas del procedimiento.

MODOS DE REALIZACIÓN DE LA INVENCIÓN

A continuación, se pasa a describir de manera breve un modo de realización de la invención, como ejemplo ilustrativo y no limitativo de esta.

El dispositivo de la realización de la figura 1 parte de unos cascos (1), que preferiblemente son del tipo con dos orejeras (11) unidas por un puente (12) ajustable. Los cascos (1) están previstos para un aislamiento elevado del ruido ambiental. Por ejemplo, pueden incluir, entre otros elementos, aros acolchados, copas protectoras y bandas ajustables para garantizar un ajuste cómodo y seguro durante su uso prolongado en situaciones de alta intensidad sonora.

Los cascos (1) poseen unos micrófonos (2) exteriores, preferiblemente cuatro en cada lado, y pudiendo disponer también en el puente (12). Esta distribución proporciona una cobertura omnidireccional completa, lo que permite una detección precisa de fuentes sonoras desde cualquier dirección. Además, estos micrófonos (2) están meticulosamente seleccionados por su capacidad para resistir condiciones climáticas adversas, incluida la exposición a la intemperie y ambientes industriales ruidosos. Su

construcción robusta garantiza un rendimiento confiable incluso en situaciones extremas. La mejor forma de lograr esta protección se consigue mediante una construcción en doble carcasa, de forma que la carcasa interior soporta los micrófonos y la exterior los protege del ambiente y de golpes, roces… Protegiéndolos de elementos externos tales como lluvia, polvo, viento y radiación UV, asegurando su correcto funcionamiento incluso en condiciones climáticas adversas.

Los micrófonos (2) tienen preferentemente una alta relación señal a ruido (mayor o igual a 60 dB). Se usan tanto para la detección como para la localización de la fuente de sonido. Dichos micrófonos serán preferentemente de tipo MEMS (con salida digital o analógica), aunque también se pueden utilizar micrófonos de tipo electret. Capturan con precisión una amplia gama de frecuencias sonoras (entre 20 Hz y 16000 Hz al menos) asegurando que ninguna señal de interés pase desapercibida. Esta capacidad de captura detallada es ventajosa para la detección precisa de sonidos de interés, como alarmas o alertas en entornos industriales, donde la seguridad del trabajador es primordial. Además, facilitan la localización precisa de la fuente de sonido al analizar diferencias en la intensidad y tiempo de llegada de la señal a cada micrófono (2).

El sistema de digitalización opera con alta tasa de muestreo y precisión para garantizar la fidelidad de las señales convertidas. Cada micrófono (2) proporciona una señal analógica única que representa las ondas sonoras capturadas, las cuales son convertidas en datos digitales mediante procesos ADC (conversión analógica-digital) de alta calidad.

También poseen unos transductores (3) interiores para reproducir al usuario los sonidos relevantes, como voces, sirenas y pitidos de alerta, etc. Un software permite realizar esa distinción y selección, procesando y filtrando la información acústica en tiempo real. Normalmente se colocarán dos transductores para cascos en cada lado del usuario, separados por una almohadilla de espuma u otro material similar.

El software se ejecuta en una unidad de control (4), a partir de la digitalización de los sonidos de los micrófonos (2). La unidad de control (4), así como los digitalizadores, se podrán alojar en una petaca o en los propios cascos (1), dependiendo de la escala de integración deseada en cada momento.

La unidad de control (4) está equipado con procesadores de alto rendimiento, memoria RAM rápida y capacidad de almacenamiento suficiente para ejecutar los algoritmos y

las inteligencias artificiales requeridas en tiempo real. Además, cuenta con interfaces de entrada y salida que permiten la comunicación con otros componentes del sistema, como los micrófonos (2) y los transductores (3).

La capacidad de procesamiento en tiempo real garantiza una respuesta instantánea del sistema ante cambios en el entorno acústico, permitiendo una detección, localización y filtrado precisos de sonidos de interés. Esta capacidad también es esencial para la implementación de algoritmos y técnicas de inteligencia artificial que mejoran la eficacia y la precisión del sistema en la identificación y clasificación de fuentes sonoras.

Una fuente de energía, como una o más baterías recargables o reemplazables permiten asegurar el funcionamiento continuo.

El software se basa en el cálculo del espectrograma de MEL de las señales digitalizadas de los micrófonos, que es tratada en una red neuronal profunda de Transformers para la inferencia de la detección del sonido. Además, una red neuronal profunda se ocupa de extraer sonidos de interés del sonido exterior.

El dispositivo identifica los sonidos de interés, como avisos acústicos, voces de otras personas que se dirigen al usuario, así como cualquier sonido relevante en la seguridad; analizando los espectrogramas previamente calculados utilizando una red Transformer aplicada al audio. Esta red está entrenada específicamente para detectar dichos sonidos de interés. Este sistema de detección se procesa en la unidad de control (4), funcionando como un disparador para las siguientes etapas del sistema.

El entrenamiento se puede realizar en el laboratorio, a partir de un sistema de reproducción de sonido 3D de alto realismo que sintetiza el campo acústico con alta precisión, y que ha sido utilizado para sintetizar las bases de datos para entrenar la Inteligencia Artificial (IA). Este sistema permite emular entornos realistas, simulando escenarios con sonido industrial y sonidos de interés provenientes de diferentes direcciones. Los datos obtenidos a partir de estas simulaciones son cruciales para el proceso de entrenamiento de la IA.

Una vez recopilados los datos de sonido, se procede a generar espectrogramas de Mel, los cuales sirven como representaciones visuales de las frecuencias presentes en los sonidos a lo largo del tiempo. Estos espectrogramas constituyen la entrada para el

modelo de IA. Para la estructura de la IA, se selecciona el modelo de Transformers, específicamente el Audio Spectrogram Transformer (AST).

Se modifica la estructura de salida del modelo AST para adaptarla a la tarea específica de detección de sonidos. La salida del modelo se ajusta para que contenga X neuronas, donde X representa el número de sonidos que se desean detectar. Cada neurona puede tomar un valor de 0 o 1, indicando la detección o no del sonido correspondiente.

Con la estructura definida, se procede al entrenamiento del modelo de IA. Durante este proceso, se proporcionan espectrogramas concretos como entradas, y se especifica cuáles de las neuronas de salida deben activarse (tomar el valor 1) y cuáles deben permanecer inactivas (valor 0). Este enfoque permite que la IA aprenda a activar las neuronas correctas en función de la entrada recibida. El proceso de entrenamiento se repite utilizando todos los datos previamente emulados con el sistema WFS, lo que garantiza que la IA adquiera la capacidad de detectar con precisión los sonidos de interés en entornos industriales simulados.

El dispositivo recoge la información de cada uno de los micrófonos (2) y extrae el espectrograma de Mel, que permite visualizar de manera clara la evolución de la trama de sonido en tiempo y frecuencia. Se utiliza una frecuencia de muestreo de, por ejemplo, 16000 Hz para capturar los sonidos de interés, suficiente para el entorno industrial y de seguridad.

La información proporcionada por el espectrograma permite una visualización detallada de la distribución espectral del sonido, lo que facilita la identificación de componentes sonoros relevantes. Además, la señal de audio original se utiliza como base para el análisis de características específicas, como la intensidad, la frecuencia y la duración de los sonidos.

Utilizando esta variedad de información, los algoritmos de extracción de sonido pueden identificar y separar de manera efectiva las señales de interés, como alarmas o discursos, del resto del sonido ambiental. Estas señales limpias y separadas se procesan luego para su posterior localización y emulación en los cascos (1) de protección auditiva.

El dispositivo recoge las muestras del último segundo de sonido y las concatena con la información que acaba de recibir, de manera que se va analizando el audio en un

contexto temporal más amplio. Es decir, el dispositivo opera utilizando una estrategia de ventana deslizante, por ejemplo, de duración 4 segundos, con desplazamiento de 1 segundo o menos, para analizar el audio en un contexto temporal más amplio.

Esta realización funciona de la siguiente manera:

1. El dispositivo recopila y almacena las muestras de sonido de un segundo.
2. Luego, recopila el siguiente segundo de sonido y lo añade al buffer.
3. Este proceso se repite hasta que el dispositivo ha recopilado un total de cuatro segundos de audio.
4. Una vez que se han acumulado cuatro segundos de audio, el dispositivo analiza estos cuatro segundos en su conjunto, proporcionando un análisis más detallado y contextualizado que si solo analizara segmentos de un segundo por separado.
5. Después del análisis, el dispositivo continúa grabando el siguiente segundo de audio.
6. Este nuevo segundo de audio se concatena con los cuatro segundos previamente almacenados, resultando en un buffer de cinco segundos.
7. Para mantener una ventana de análisis constante de cuatro segundos, el dispositivo descarta el primer segundo de audio, el más antiguo, y conserva los últimos cuatro segundos (los tres segundos más recientes y el nuevo segundo añadido).
8. Este proceso de recopilación, concatenación y descarte se repite continuamente, asegurando que siempre se analicen los últimos cuatro segundos de audio.

Esta estrategia permite al dispositivo mantener un contexto temporal amplio y actualizado del sonido ambiente, mejorando la precisión del análisis y detección de sonidos relevantes.

Un detector identifica los sonidos de interés, como avisos acústicos, voces de otras personas que se dirigen al usuario, así como cualquier sonido relevante en la seguridad; analizando los espectrogramas previamente calculados utilizando una red Transformer aplicada al audio. Esta red está entrenada específicamente para detectar esos sonidos de interés. Este sistema de detección se procesa en la unidad de control (4), funcionando como un disparador para las siguientes etapas del sistema. La inteligencia artificial de detección está entrenada para identificar sonidos específicos como avisos acústicos, voces y sonidos relevantes para la seguridad, utilizando una red Transformer aplicada a espectrogramas de audio. La potencia del Transformer radica en su capacidad para discernir entre señales de interés y ruido de fondo, basándose en patrones complejos y contextuales aprendidos durante su entrenamiento.

En lugar de limitarse a rangos específicos de frecuencia o amplitud, el Transformer considera características adicionales como la cadencia y el ritmo de los sonidos, entre otros muchos patrones extraíbles. Por ejemplo, aunque la voz humana generalmente se encuentra en ciertos rangos de frecuencia, la red también tiene en cuenta estos patrones temporales para mejorar su precisión en la identificación de sonidos relevantes.

Además, los catalogados como “sonidos de interés” pueden modificarse y ampliarse mediante un nuevo entrenamiento de la red, permitiendo que se incorporen nuevos tipos de audio de interés según las necesidades específicas. Durante este proceso, se pueden etiquetar los audios relevantes para que la red aprenda a reconocerlos con precisión.

La invención incluye tres configuraciones o realizaciones preferidas de micrófonos instalados en los cascos. Una primera con ocho micrófonos (2), situándose cuatro en cada orejera, y una segunda con cuatro micrófonos (2), situándose uno en cada orejera y dos en la parte superior del puente. Una tercera, que requiere menos capacidad de procesamiento, posee un micrófono (2) en cada orejera y dos en el puente. En general se prefiere que no haya menos de cuatro micrófonos (2).

Para la localización precisa de eventos sonoros de interés, se utiliza la señal grabada y limpiada por cada uno de los micrófonos (2) del sistema. El método se basa en el cálculo del GCC-PHAT (Correlación Cruzada Generalizada con corrección de fase (GCC-PHAT)) de los distintos pares de micrófonos (2). Así, se combina la señal de cada micrófono (2) individual con la señal de cada otro micrófono (2) (28 pares en el caso de utilizar 8 micrófonos, y 6 pares en el caso de utilizar 4).

La característica GCC-PHAT (Generalized Cross-Correlation with Phase Transform) es una métrica que se obtiene al analizar la correlación entre señales de audio captadas por distintos micrófonos (2), evaluando el tiempo de retardo con el que el sonido llega a cada uno de ellos. Este análisis produce una señal GCC-PHAT cuya longitud es el doble de la longitud de las señales originales, con el punto central representando un desfase de 0. Si la señal GCC-PHAT se desplaza hacia un lado del punto central, indica que una señal se adelanta respecto a la otra, y viceversa.

Dependiendo de dónde se concentre la energía en la señal GCC-PHAT, se puede determinar el retardo con el que la señal ha llegado a cada uno de los micrófonos.

Dado que la disposición física de los micrófonos (2) determina un límite máximo para los retardos posibles (en función de la distancia entre ellos), no es necesario utilizar toda la señal GCC-PHAT resultante para la inferencia. Podemos optimizar el proceso recortando la señal GCC-PHAT únicamente a la región relevante, que corresponde al intervalo de retardos máximos posibles entre los micrófonos. Este recorte reduce la carga computacional al disminuir la cantidad de datos que deben ser procesados, sin perder información esencial para la inferencia de la red neuronal.

Este proceso proporciona información crucial sobre la diferencia de tiempo de llegada de la señal a cada par de micrófonos, lo que ayuda a determinar la dirección del sonido.

Una vez calculada la GCC-PHAT para cada par de micrófonos (2), se procede a concatenar todas las muestras GCC-PHAT posibles en la tercera dimensión, similar a una imagen RGB que tiene los canales correspondientes al rojo, verde y azul. Esto permite tener toda la información de correlación cruzada de todos los micrófonos posibles para un mismo sonido en una estructura de datos tridimensional.

Esta información, que contiene las características de la distribución espacial del sonido capturado por los micrófonos, se utiliza para alimentar una red neuronal profunda capaz de aprender y extraer patrones complejos de la información de la GCC-PHAT, permitiendo así determinar la dirección del sonido de donde proviene el evento sonoro de interés, ya sea una alarma o un discurso.

En el proceso de localización de eventos sonoros, se emplean redes neuronales avanzadas, como las ResNet (Redes Convolucionales Residuales), para realizar inferencias precisas sobre la dirección de los eventos en función de la información proporcionada por el GCC-PHAT. Las ResNet son conocidas por su capacidad para aprender representaciones profundas y jerárquicas de los datos, lo que las hace ideales para tareas de procesamiento de señales como la localización de eventos sonoros.

Estas redes se alimentan con la información tridimensional del GCC-PHAT, que contiene la correlación cruzada de las señales de audio entre todos los pares de micrófonos (2) posibles. Las capas iniciales de la ResNet procesan esta información, extrayendo características relevantes y aprendiendo representaciones profundas de los datos.

A medida que la información avanza a través de las capas de la ResNet, la red aprende a identificar patrones complejos en los datos de GCC-PHAT que están asociados con la dirección de los eventos sonoros de interés. Utilizando estas características aprendidas, la red puede realizar inferencias precisas sobre la ubicación espacial de los eventos sonoros en función de la información proporcionada por los micrófonos.

Una vez completado el proceso, los grados detectados por el sistema se pasan al siguiente paso del sistema, donde se hace uso de estos grados para representar de manera real de dónde viene cada sonido. Esta IA es capaz de proporcionar un ángulo en azimuth y otro en elevación con respecto al frente de los cascos (1), lo que indica la dirección desde la cual proviene el sonido de interés. Esta información es fundamental para el proceso de representación espacial del sonido dentro del sistema, permitiendo una emulación precisa de la ubicación del evento sonoro en el entorno del usuario.

Por otro lado, la emisión de sonido se puede personalizar a partir de una HRTF (Función de Transferencia de Respuesta al Cabeza) que permite adaptar la emisión de los auriculares para que el usuario reconozca con precisión el origen del sonido. Estas funciones se individualizan en una red neuronal profunda a partir de imágenes de las orejas y la cabeza del usuario. Luego esas funciones se aplican mediante un algoritmo que permite representar los sonidos en los transductores (3) de forma que el usuario identifica la dirección desde donde llegan.

Este proceso se lleva a cabo previamente al primer uso de los cascos por parte del usuario. El objetivo principal es proporcionar una experiencia auditiva personalizada al usuario, adaptando la respuesta de los transductores (3) a las características fisiológicas únicas de sus orejas, cabeza y torso.

Para lograr este objetivo, se utilizará una HRTF individualizada. La HRTF es el filtro que describe cómo un humano recibe el sonido desde las diferentes direcciones del espacio. Es única para cada oyente y está fuertemente relacionada con las estructuras morfológicas de cada oyente, como la cabeza, los hombros y principalmente la forma de las orejas. Debido a su singularidad, el uso de una HRTF genérica para aplicaciones binaurales a menudo genera errores de localización y una limitada percepción espacial.

Para el proceso de individualización de la HRTF, se utiliza un modelo de inteligencia artificial *end-to-end*, el cual, mediante imágenes RGB de la cabeza, hombros y orejas encuentra patrones antropométricos que luego utiliza para inferir la HRTF perteneciente

a cada individuo. El modelo end-to-end consta de tres partes fundamentales. Primeramente, un modelo profundo del tipo YOLO (YOU ONLY LOOK ONCE) para detectar orejas humanas, luego de la detección se procede a recortar y extraer las imágenes de las orejas desde diferentes ángulos. Las imágenes de las orejas previamente recortadas alimentan un modelo profundo que extrae complejos patrones antropométricos, mediante el uso de capas convolucionales y mecanismos de atención. Finalmente, se emplea un modelo que decodificará el vector de características antropométricas en la HRTF de cada individuo.

La HRTF individualizada se adapta específicamente a las características anatómicas únicas del usuario, pudiendo incluir auditivas (frecuencias que oye mejor), teniendo en cuenta factores como la dirección del sonido, la distancia y la posición relativa de las fuentes sonoras. Esto garantiza una representación precisa y realista del sonido en los transductores del usuario, lo que mejora significativamente la inmersión y la calidad de la experiencia auditiva.

En la etapa final del sistema, se emplea toda la información recopilada y procesada previamente para ofrecer al usuario una experiencia auditiva inmersiva y altamente realista. Con la detección precisa de sonidos por parte del modelo de Transformers, la obtención de audios de interés completamente libres de ruido ambiente, la localización precisa mediante el sistema de IA y el modelo de HRTF individualizado del usuario, el algoritmo finaliza la tarea de representar sonidos localizados a través de la HRTF personalizada.

La idea principal detrás de este algoritmo es reproducir el sonido de interés a través de los transductores (3) integrados en los cascos (1). Para lograr esto, se aprovecha la información obtenida del proceso de localización, que proporciona la dirección desde la cual se detectó el sonido. Utilizando la HRTF individualizada del usuario, el algoritmo aplica las características anatómicas específicas del usuario para emular con precisión cómo el sonido llegaría a sus oídos desde esa dirección particular en el espacio tridimensional.

Al hacer uso de la HRTF personalizada, el algoritmo garantiza que el sonido reproducido a través de los transductores coincida de manera precisa con la percepción auditiva del usuario en su entorno real. Esto significa que el usuario experimentará el sonido de interés como si estuviera realmente presente en la dirección desde la cual fue detectado,

con una claridad y una fidelidad excepcionales y sin ningún tipo de interferencia del ruido ambiental.

Los cascos (1) pueden tener un receptor inalámbrico para emitir sonidos directamente enviados directamente a ellos. Así, un centro de control puede enviar señales a los cascos (1) de un trabajador concreto, de un grupo de trabajadores o a todos.

REIVINDICACIONES

1- Dispositivo de protección auditiva, formado por unos cascos (1), unos micrófonos (2) exteriores y unos transductores (3) interiores, gestionados por una unidad de control (4), caracterizado por que los cascos (1) son de protección auditiva pasiva, configurados para eliminar el sonido ambiental, y la unidad de control (4) está configurada para seleccionar las señales relevantes del sonido captado por los micrófonos (2) y emitirlas por los transductores (3).

2- Dispositivo de protección auditiva, según la reivindicación 1, caracterizado por que la unidad de control (4) está configurada para adaptar la señal emitida al usuario para simular la proveniencia del sonido relevante.

3- Dispositivo de protección auditiva, según la reivindicación 1, caracterizado por que comprende cuatro micrófonos (2) en cada lado, o un micrófono en cada lado junto a dos micrófonos en la diadema

4- Procedimiento de protección auditiva, en el dispositivo de la reivindicación 1, caracterizado por que comprende las etapas de:
Bloquear los sonidos ambientales mediante unos cascos (1);
Captar los sonidos ambientales mediante unos micrófonos (2);
Extraer unos sonidos relevantes de los sonidos ambientales;
Emitir los sonidos ambientales en unos transductores (3) interiores a los cascos (1).

5- Procedimiento de protección auditiva, según la reivindicación 4, caracterizado por que comprende una etapa de adaptación de la emisión de los transductores (3) a la fisiología del usuario.

Fig. 1

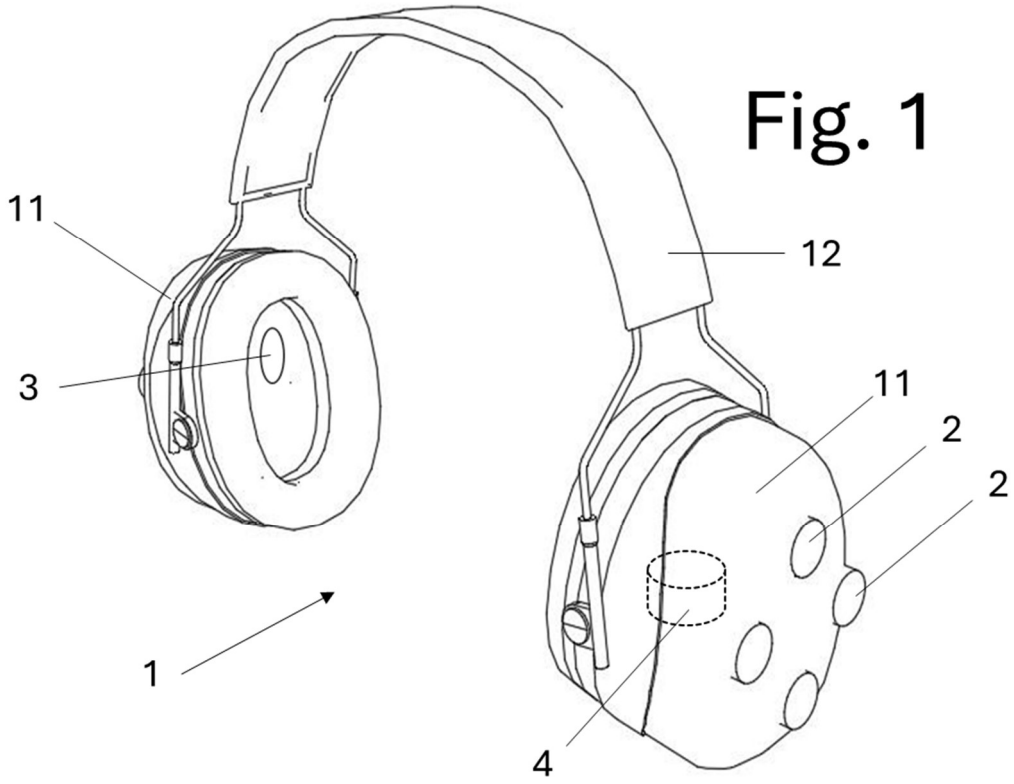
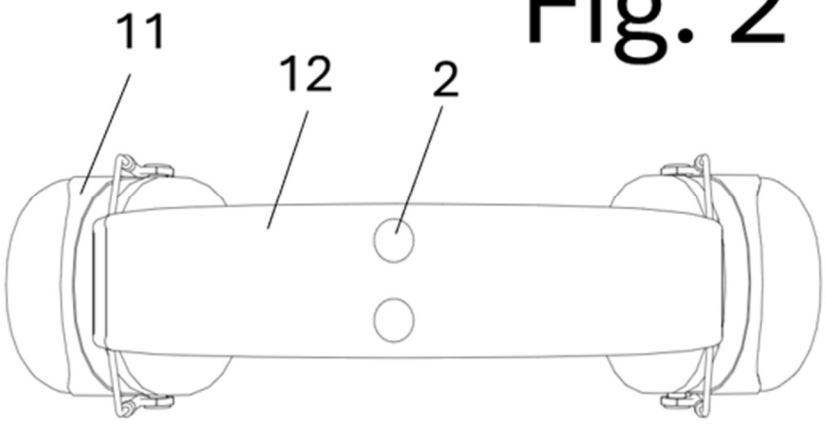


Fig. 2



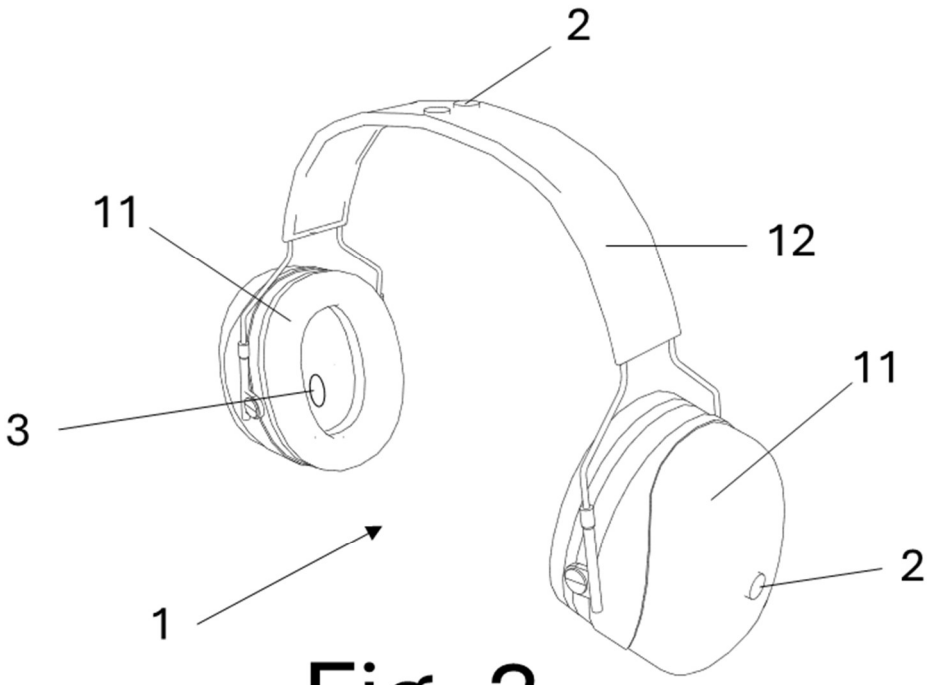
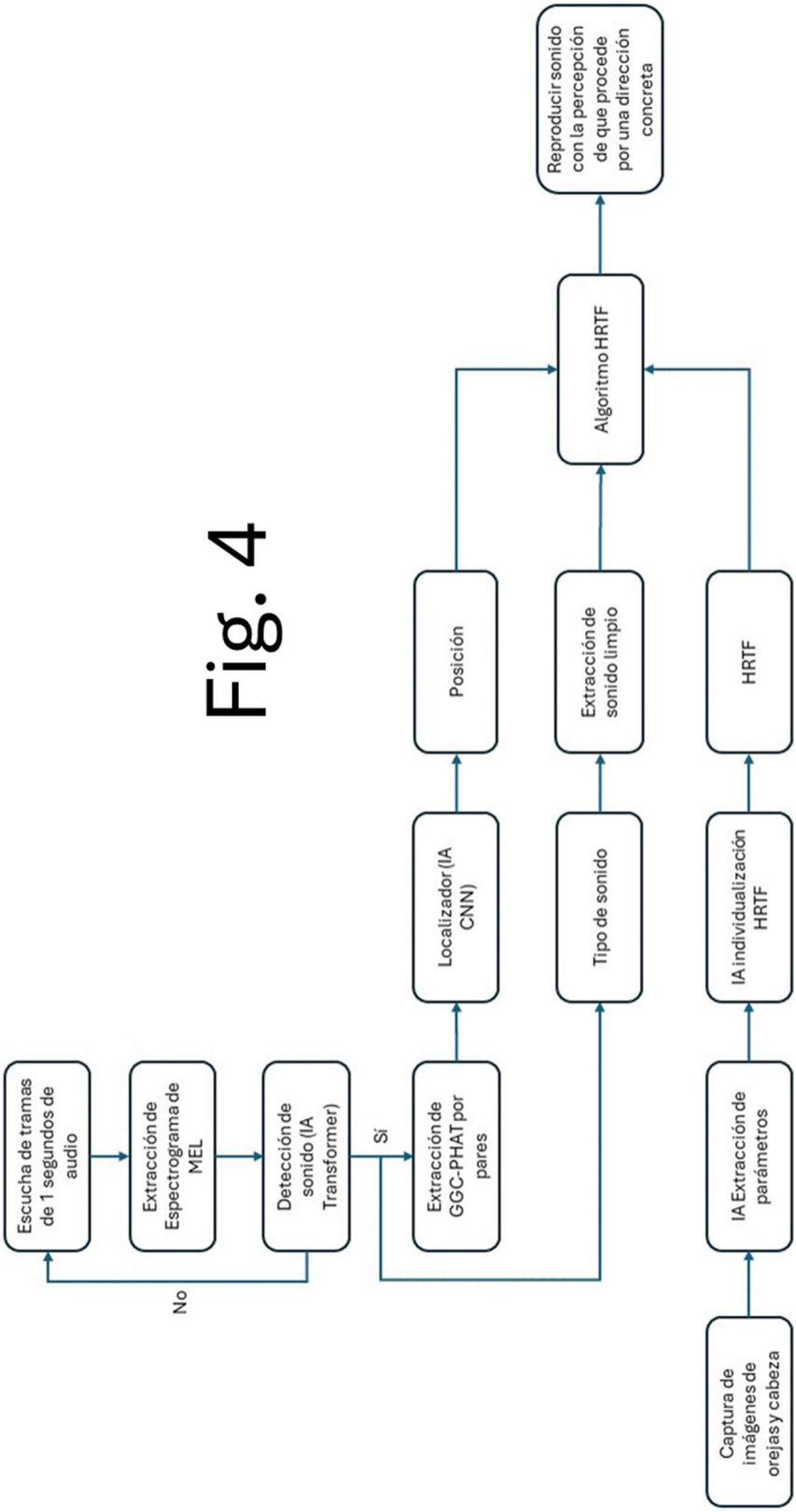


Fig. 3

Fig. 4


Escucha de tramas de 1 segundos de audio
Extracción de Espectrograma de MEL
Detección de sonido (IA Transformer)
No
Sí
Extracción de GGC-PHAT por pares
Localizador (IA CNN)
Posición
Tipo de sonido
Extracción de sonido limpio
Captura de imágenes de orejas y cabeza
IA Extracción de parámetros
IA individualización HRTF
HRTF
Algoritmo HRTF
Reproducir sonido con la percepción de que procede por una dirección concreta

OFICINA ESPAÑOLA
DE PATENTES Y MARCAS

ESPAÑA

(21) N.º solicitud: 202430765

(22) Fecha de presentación de la solicitud: 26.09.2024

(32) Fecha de prioridad:

INFORME SOBRE EL ESTADO DE LA TECNICA

(51) Int. **CI.:** Ver Hoja Adicional

DOCUMENTOS RELEVANTES

Categoría	(56) **Documentos citados**	**Reivindicaciones afectadas**
X Y	VELURI, BANDHAV et al.: "Semantic hearing: Programming acoustic scenes with binaural hearables". Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 2023, Páginas 1-15.	1-4 5
Y	YANG, ZHIJIAN; CHOUDHURY, ROMIT ROY: "Personalizing head related transfer functions for earables". Proceedings of the 2021 ACM SIGCOMM 2021 Conference, 2021, Páginas 137-150.	5
X	US 2015294662 A1 (IBRAHIM) 15/10/2015, página 2, párrafo [21] - página 3, párrafo [24]; página 3, párrafo [29] - página 4, párrafo [31]; figuras 1, 6.	1, 3
A	XIAO, TONG; XU, BUYE; ZHAO, CHUMING: "Spatially selective active noise control systems". The Journal of the Acoustical Society of America, 2023, Vol. 153, Nº 5, Páginas 2733-2734.	1-4
A	JAYARAM, VIVEK; KEMELMACHER-SHLIZERMAN, IRA; SEITZ, STEVEN M.: "HRTF Estimation in the Wild". Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 2023, Páginas 1-9.	1, 2

Categoría de los documentos citados

X: de particular relevancia
Y: de particular relevancia combinado con otro/s de la misma categoría
A: refleja el estado de la técnica
O: referido a divulgación no escrita
P: publicado entre la fecha de prioridad y la de presentación de la solicitud
E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

El presente informe ha sido realizado

☒ para todas las reivindicaciones ☐ para las reivindicaciones nº:

Fecha de realización del informe	**Examinador**	**Página**
22.10.2024	R. San Vicente Domingo	1/2

CLASIFICACIÓN OBJETO DE LA SOLICITUD

A61F11/14 (2006.01)
G10K11/16 (2006.01)
H04R1/10 (2006.01)

Documentación mínima buscada (sistema de clasificación seguido de los símbolos de clasificación)

A61F, G10K, H04R

Bases de datos electrónicas consultadas durante la búsqueda (nombre de la base de datos y, si es posible, términos de búsqueda utilizados)

INVENES, EPODOC