(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 619 615**

(51) Int. CI.:

G10L 15/24 (2013.01)
G06F 3/01 (2006.01)
G06K 9/00 (2006.01)
G06K 9/62 (2006.01)
A63F 13/424 (2014.01)
A63F 13/213 (2014.01)
G10L 15/22 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **18.10.2013 PCT/US2013/065793**

(87) Fecha y número de publicación internacional: **01.05.2014 WO2014066192**

(96) Fecha de presentación y número de la solicitud europea: **18.10.2013 E 13783214 (3)**

(97) Fecha y número de publicación de la concesión europea: **21.12.2016 EP 2912659**

(54) Título: **Aumento del reconocimiento de voz con imágenes de profundidad**

(30) Prioridad:

26.10.2012 US 201213662293

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
26.06.2017

(73) Titular/es:

MICROSOFT TECHNOLOGY LICENSING, LLC (100.0%)
One Microsoft Way
Redmond, WA 98052, US

(72) Inventor/es:

KAPUR, JAY;
TASHEV, IVAN;
SELTZER, MIKE y
HODGES, STEPHEN, EDWARD

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Aumento del reconocimiento de voz con imágenes de profundidad

Antecedentes

El reconocimiento de voz informatizado busca identificar palabras pronunciadas desde información de audio, tal como desde señales de audio recibidas a través de uno o más micrófonos. Un ejemplo de reconocimiento de voz informatizado se describe en el documento US 2009/0138805 A1. Sin embargo, pueden surgir ambigüedades en la identificación de palabras pronunciadas en la información de audio. Además, el contexto de las palabras pronunciadas, por ejemplo, donde las palabras pronunciadas estaban destinadas a ser una entrada de voz a un dispositivo de ordenador, puede no ser fácilmente determinado desde tal información de audio.

Resumen

Se describen realizaciones relativas al uso de imágenes de profundidad para aumentar el reconocimiento de voz. Por ejemplo, una realización descrita proporciona, en un dispositivo informático, un método que incluye la recepción de información de profundidad de un espacio físico desde una cámara de profundidad, recibiendo información de audio desde uno o más micrófonos, identificando un conjunto de una o más posibles palabras pronunciadas desde la información de audio, determinando una entrada de voz para el dispositivo informático basada en la comparación del conjunto de una o más posibles palabras pronunciadas desde la información de audio y la información de profundidad, y realizando una acción en el dispositivo informático basada en la entrada de voz determinada.

Se proporciona este resumen para introducir una selección de conceptos de una forma simplificada que además son descritos a continuación en la Descripción Detallada, mientras que la invención es definida por una reivindicación 1 de método, independiente y por una reivindicación 11 de dispositivo, independiente.

Además, el objeto en cuestión reivindicado no está limitado a las implementaciones que solucionan alguno o todos los inconvenientes señalados en alguna parte de esta descripción.

Breve descripción de los dibujos

La FIG. 1 muestra un ejemplo esquemático de un entorno de reconocimiento de voz según una realización de la descripción.

La FIG. 2 es un diagrama de flujo que ilustra un método para el reconocimiento de voz según una realización de la descripción.

La FIG. 3 es un diagrama de flujo que ilustra un método para el reconocimiento de voz según otra realización de la descripción.

La FIG. 4 es un diagrama de flujo que ilustra un método para el reconocimiento de voz según otra realización más de la descripción.

La FIG. 5 muestra esquemáticamente un sistema informático no limitativo.

Descripción detallada

El reconocimiento de voz informatizado puede plantear varios desafíos. Por ejemplo, la pronunciación de palabras individuales, el acento, la nitidez, el tono, las imperfecciones/impedimentos, y otras variables de la voz humana que pueden diferir ampliamente entre usuarios. Adicionalmente, la reverberación y/o el ruido y otros sonidos no deseados (p.ej., altavoces, aspiradoras, etc.) en la habitación en la que las palabras son pronunciadas pueden dificultar el reconocimiento de voz. Además, el contexto en el que las palabras reconocidas son pronunciadas puede afectar a factores tales como si un segmento de voz reconocida estuviera destinada como una entrada de voz.

Por consiguiente, se han descrito realizaciones relativas a aumentar un proceso de reconocimiento de voz con información literal y/o contextual identificada en la información de profundidad recibida desde una cámara de profundidad. Por ejemplo, en algunas realizaciones, los movimientos de la boca, la lengua, y/o la garganta del orador pueden ser identificados desde la información de profundidad y utilizados para confirmar la identidad de posibles palabras pronunciadas identificadas a través de datos de audio, identificar palabras no detectadas en los datos de audio, etc. Adicionalmente, en algunas realizaciones, gestos, posturas, etc., realizados por el orador pueden ser identificados desde la información de profundidad y utilizados para colocar las palabras identificadas en un contexto deseado, tal como la confirmación de que las palabras pronunciadas identificadas tenían por objeto ser una entrada a un dispositivo informático. El término “reconocimiento de voz” como se ha usado aquí puede incluir reconocimiento de palabras, reconocimiento del orador (p.ej. cuál de dos o más oradores está hablando en un entorno), reconocimiento semántico, reconocimiento de las emociones, y/o el reconocimiento de cualquier otro aspecto adecuado de la voz en un entorno de uso.

La FIG. 1 muestra un ejemplo no limitativo de un entorno 100 de reconocimiento de voz. En particular, la FIG. 1 muestra un sistema informático 102 en la forma de una consola de entretenimiento que puede ser usada para jugar a una variedad diferente de juegos, reproducir uno o más tipos diferentes de medios, y/o controlar o manejar aplicaciones que no son juegos y/o sistemas operativos. La FIG. 1 muestra también un dispositivo 104 de visualización tal como una televisión o un monitor de ordenador, que puede ser usado para mostrar a los usuarios contenido de medios, imágenes de juego, contenido informático que no es de juegos, etc.

El entorno 100 de reconocimiento de voz incluye además un dispositivo 106 de captura en forma de una cámara de profundidad que monitoriza visualmente o sigue objetos y usuarios dentro de una escena observada. El dispositivo 106 de captura puede estar conectado operativamente al sistema informático 102 a través de uno o más interfaces. Como un ejemplo no limitativo, el sistema informático 102 puede incluir un bus serie universal al que el dispositivo 106 de captura puede estar conectado. El dispositivo 106 de captura puede ser utilizado para reconocer, analizar, y/o seguir uno o más sujetos humanos y/o objetos dentro del espacio físico, tal como un usuario 108. En un ejemplo no limitativo, el dispositivo 106 de captura puede incluir una fuente de luz infrarroja para proyectar luz infrarroja sobre el espacio físico y una cámara de profundidad configurada para recibir luz infrarroja. El dispositivo de captura puede también incluir otros sensores, incluyendo pero sin limitarse a un sensor o sensores de imagen bidimensional (p.ej. una cámara de luz visible tal como un sensor de imagen RGB y/o un sensor de escala de grises) y uno o más micrófonos (p.ej. una agrupación de micrófonos direccionales). Aunque se describe como que proporciona una entrada a una consola de entretenimiento, se entenderá que una cámara de profundidad puede ser utilizada para proporcionar una entrada relevante para reconocimiento de voz para cualquier sistema informático adecuado, y puede ser utilizada en entornos que no sean de juego.

Con el fin de formar imágenes de objetos dentro del espacio físico, la fuente de luz infrarroja puede emitir luz infrarroja que se refleja en los objetos en el espacio físico y que es recibida por la cámara de profundidad. Basándose en la luz infrarroja recibida, se puede construir un mapa de profundidad del espacio físico. El dispositivo 106 de captura puede emitir el mapa de profundidad obtenido desde la luz infrarroja al sistema informático 102, donde puede ser utilizado para crear una representación del espacio físico cuya imagen se ha formado por la cámara de profundidad. El dispositivo de captura puede también ser utilizado para reconocer objetos en el espacio físico, monitorizar el movimiento de uno o más usuarios, realizar el reconocimiento de gestos, etc. Cualquier tecnología de búsqueda de profundidad puede ser virtualmente utilizada sin apartarse del alcance de esta descripción. Ejemplos de tecnologías de búsqueda de profundidad se explican con más detalle con referencia a la FIG. 5.

La FIG. 1 también muestra un escenario en el que el dispositivo 106 de captura sigue al usuario 108 de forma que los movimientos del usuario pueden ser interpretados por el sistema informático 102. En particular, los movimientos de la boca, la lengua, y/o la garganta del usuario 108 pueden ser monitorizados para determinar si el usuario 108 está hablando. Si el usuario 108 está hablando, la información de audio recibida por el sistema informático 102 (p.ej. a través de uno o más micrófonos incorporados en el dispositivo 106 de captura y/o situados fuera del dispositivo 106 de captura) puede ser analizada para reconocer una o más de las palabras pronunciadas por el usuario. Los movimientos de la boca, la lengua y/o la garganta pueden ser también utilizados para aumentar el proceso de identificación de las palabras pronunciadas, por ejemplo, confirmando que las palabras identificadas fueron pronunciadas, añadiendo palabras pronunciadas adicionales identificadas.

La información procedente del dispositivo de captura puede también ser utilizada para determinar varios elementos contextuales de las palabras pronunciadas identificadas. Por ejemplo, si usuarios adicionales están presentes en el espacio físico, tal como el usuario 110, el usuario desde el que las palabras pronunciadas fueron recibidas puede distinguirse de otros usuarios comparando las palabras pronunciadas con los movimientos de la boca/lengua/garganta de uno o más usuarios en el espacio físico. Además, el reconocimiento facial, identificación del orador (p.ej. basado en la altura, forma del cuerpo, modo de andar, etc. del usuario), y/u otras técnicas válidas adicionales pueden ser utilizadas para determinar la identidad de la persona que está hablando. Las posiciones y/u orientaciones relativas de uno o más usuarios en una habitación se pueden también seguir para ayudar a determinar si un orador está haciendo una entrada de voz. Por ejemplo, si un usuario no está mirando al dispositivo de captura cuando está hablando, se puede determinar que el usuario no está hablando al sistema. Igualmente, donde múltiples usuarios son visibles por el dispositivo de captura, que un usuario esté mirando al dispositivo de captura puede ser utilizado como información para identificar qué persona hizo una entrada de voz.

Además, una vez que uno o más usuarios han sido identificados, se puede hacer el seguimiento del uno o más usuarios (a través del dispositivo de captura, por ejemplo). Esto puede ayudar a facilitar la correspondencia eficaz de futuras voces reconocidas a oradores identificados, y por lo tanto identificar rápidamente qué modelo/parámetros de reconocimiento de voz usar para un usuario en particular (p.ej. para afinar el reconocimiento de voz de ese usuario).

Además, los gestos realizados por el usuario 108 identificados a través de la información procedente del dispositivo 106 de captura pueden ser utilizados para identificar información contextual relativa a palabras pronunciadas identificadas. Por ejemplo, si el usuario 108 está hablando con la intención de controlar el sistema informático 102 a través de comandos de voz, el usuario 108 puede realizar uno o más gestos y/o posturas, deliberadamente o no, que pueden indicar esta intención. Ejemplos incluyen, pero no se limitan a, apuntar hacia el dispositivo 104 de

visualización, mirando al sistema informático 102 o dispositivo 104 de visualización mientras se habla, o realizar un gesto específico que está asociado con una entrada de usuario reconocida. Así, mediante la identificación del gesto realizado por el usuario 108 así como la identificación de las palabras pronunciadas, se puede determinar la intención del usuario para controlar el dispositivo del ordenador. Igualmente, si el usuario 108 está mirando a otro usuario, gesticulando a otro usuario, etc., mientras está hablando, en algunas realizaciones no puede deducirse una intención de controlar el dispositivo informático.

Otros tipos de información contextual pueden ser determinados igualmente desde la información recibida desde dispositivo 106 de captura. Por ejemplo, en algunas realizaciones, se puede determinar un estado emocional del usuario 108 cuando está hablando por los rasgos faciales y/o corporales, posturas, gestos, etc. del usuario 108 desde la información de profundidad. Como otro ejemplo aún, los objetos en el espacio físico visualizado pueden ser identificados y utilizados para distinguir palabras ambiguas. Por ejemplo, puede ser difícil distinguir palabras compuestas como “quarterback” de las palabras individuales (“quarter” y “back”) que configuran la palabra compuesta. Por lo tanto, en el caso de tales ambigüedades, los datos de imagen de profundidad del espacio físico pueden ser utilizados para detectar objetos, acciones, etc., que pueden proporcionar el contexto para ayudar a determinar la palabra o palabras reales pronunciadas. En el ejemplo específico de “quarterback” los datos de imagen de profundidad pueden ser analizados para determinar la presencia de objetos y/u otras pistas contextuales para ayudar a eliminar la ambigüedad de estos términos, tal como dinero en una mano del usuario, objetos relativos al fútbol americano (p.ej. si el usuario está sentado delante del televisor viendo un partido de fútbol americano), etc. Tal información puede también ser utilizada en algunos ejemplos para ayudar a la eliminación de ambigüedad en los homónimos, tal como “ate” (comió) y “eight” (ocho).

El sistema informático 102 puede también ser configurado para comunicarse con uno o más dispositivos informáticos remotos, no mostrados en la FIG. 1. Por ejemplo, el sistema informático 102 puede recibir contenido de vídeo directamente desde una emisora, servicios de entrega de medios de terceros, u otro proveedor de contenido. El sistema informático 102 puede también comunicarse con uno o más servicios remotos a través de Internet u otra red, por ejemplo con el fin de analizar el audio y/o los datos de imagen recibidos, realizar el reconocimiento de voz, etc. Mientras que la realización representada en la FIG. 1 muestra el sistema informático 102, el dispositivo 104 de visualización, y el dispositivo 106 de captura como elementos separados, en algunas realizaciones uno o más de los elementos pueden estar integrados en un dispositivo común.

La FIG. 2 muestra un diagrama de flujo que representa una realización de un método 200 para el reconocimiento de voz de un usuario. El método 200 puede ser realizado por un dispositivo informático configurado para recibir y procesar audio e información de profundidad, tal como la información recibida desde el dispositivo 106 de captura.

En 202, el método 200 incluye la recepción de información de profundidad desde una cámara de profundidad. Como se ha explicado anteriormente, la información de profundidad puede ser utilizada para construir un mapa de profundidad del espacio físico cuya imagen se ha formado incluyendo uno o más usuarios. Adicionalmente, se puede también recibir información de imagen desde una cámara de luz visible. En 204, el método 200 incluye la recepción de información de audio recibida a través de uno o más micrófonos, que en algunas realizaciones puede incluir micrófonos direccionales. En 206, una o más posibles palabras pronunciadas son identificadas desde la información de audio. La una o más posibles palabras pronunciadas pueden ser identificadas por el dispositivo informático utilizando cualquier proceso de reconocimiento de voz adecuado.

En 208, el método 200 incluye la determinación de una entrada de voz al dispositivo de ordenador basada en una o más posibles palabras pronunciadas y la información de profundidad. La entrada de voz puede incluir un comando que indica que una acción ha de ser realizada por el dispositivo de ordenador, contenido destinado a ser visualizado en un dispositivo de visualización y/o grabado por un dispositivo informático, y/o cualquier otra entrada de voz adecuada.

Las posibles palabras pronunciadas identificadas y la información de profundidad pueden ser utilizadas de cualquier manera adecuada para determinar la entrada de voz. Por ejemplo, como se indica en 210, los movimientos de la boca, la lengua y/o la garganta del usuario pueden ser utilizados para determinar posibles sonidos y/o palabras pronunciados por el usuario. Estos posibles sonidos/palabras identificados pueden entonces ser utilizados para eliminar la ambigüedad de cualquier posible palabra pronunciada potencialmente ambigua desde la información de audio, y/o para incrementar una certeza de identificaciones de palabras, como se describe a continuación en más detalle.

De forma similar, en algunas realizaciones, los movimientos de la boca, la lengua y/o la garganta pueden ser utilizados para determinar independientemente un conjunto de posibles palabras pronunciadas. Este conjunto de posibles palabras pronunciadas puede ser igualmente comparado con el conjunto de posibles palabras pronunciadas determinadas desde la información de audio para ayudar a eliminar la ambigüedad de cualquier incertidumbre en la correcta identificación de palabras desde la información de audio, para añadir cualesquiera palabras potenciales omitidas a los datos de audio, etc.

Como se ha mencionado anteriormente, la información de profundidad puede también ser utilizada para identificar los elementos contextuales relativos a los posibles segmentos del discurso, como se indica en 212. Cualquier

elemento contextual adecuado puede ser identificado. Ejemplos de tales elementos contextuales pueden incluir, pero no limitarse a, una identidad del usuario, una emoción del usuario, un gesto realizado por el usuario, uno o más objetos físicos en el espacio físico del usuario, etc. Los elementos contextuales identificados desde la información de profundidad pueden ser utilizados para confirmar una entrada de voz identificada desde la información de audio, eliminar cualquier posible ambigüedad de las palabras pronunciadas (p.ej. palabras compuestas, homónimos, etc.), ubicar la entrada de voz en el contexto deseado, utilizar un sistema de micrófono direccional para aislar al orador de los demás en el entorno, afinar el reconocimiento de voz basado en los atributos conocidos del discurso del usuario identificado, y/o para cualesquiera otros propósitos pertinentes.

Continuando con la FIG. 2, el método 200 incluye, en 214, emprender una acción en el dispositivo informático basándose en la entrada de voz. Por ejemplo, se puede realizar una acción indicada por un comando de entrada de voz, se puede visualizar en el dispositivo de visualización contenido de texto correspondiente a las palabras pronunciadas, etc. Además, en algunas realizaciones, el contenido de texto puede ser etiquetado con un estado emocional, de tal modo que esas palabras pueden tener una apariencia diferente dependiendo del estado emocional detectado en el usuario cuando pronuncia las palabras.

La FIG. 3 muestra un diagrama de flujo que representa una realización de un método 300 para el reconocimiento de un comando de entrada de voz configurado para causar que un dispositivo de ordenador realice una acción específica. El método 300 puede ser realizado por un dispositivo informático configurado para recibir y procesar entrada de audio y de profundidad. En 302, el método 300 incluye la recepción de información de profundidad desde la cámara de profundidad, y en 304, la recepción de información de audio desde uno o más micrófonos. En 306, el método 300 incluye la identificación de una o más posibles palabras pronunciadas de la información de audio, y en 308, la identificación de elementos contextuales desde la información de profundidad. Los elementos contextuales pueden incluir, pero no limitarse a, un gesto realizado por el usuario (p.ej. movimiento de la boca, la garganta, la lengua, el cuerpo, etc.), como se indica en 310, un estado físico de un usuario (p.ej. si un usuario está sentado, agachado o de pie, si la boca de un usuario está abierta o cerrada, cómo de lejos está el usuario de la pantalla, una orientación de la cabeza del usuario, etc,) como se indica en 312, y/o un estado emocional del usuario, como se indica en 314. Se entenderá que estos elementos contextuales se describen con el propósito de ejemplo, y no están de ningún modo destinados ser una limitación.

En 316, el método 300 incluye una comparación de las palabras pronunciadas y los elementos contextuales identificados. Las palabras pronunciadas y los elementos contextuales pueden ser comparados para determinar, por ejemplo, si las palabras pronunciadas están destinadas como entrada de voz dirigida al dispositivo informático para realizar una acción específica basada en uno o más de los elementos contextuales identificados desde la información de profundidad. Por ejemplo, un gesto particular realizado por el usuario e identificado desde la información en profundidad puede indicar que las palabras pronunciadas pretenden ser entrada de usuario. Como un ejemplo más específico, el usuario puede dirigir un gesto al dispositivo del sistema de reconocimiento de voz, tal como es apuntar al dispositivo de ordenador/pantalla/dispositivo de captura/etc., mientras habla, y/o el usuario puede realizar un gesto que se corresponde con un gesto conocido asociado con la entrada de usuario.

Además, una orientación de la cabeza del usuario puede ser usada para determinar si las palabras pronunciadas pretenden ser una entrada de usuario. Por ejemplo, si el usuario está mirando en una dirección particular mientras está hablando, tal como hacia el dispositivo del sistema de reconocimiento de voz (p.ej. un dispositivo de visualización, un dispositivo de ordenador, un dispositivo de captura, etc.), se puede determinar que las palabras pretenden ser una entrada de usuario al dispositivo informático. Asimismo, si el usuario está mirando a otro usuario en el espacio físico mientras está hablando, se puede indicar que las palabras no pretenden ser una entrada de usuario.

En un ejemplo más, se pueden determinar una o más emociones del usuario desde los datos de profundidad para determinar si las palabras pronunciadas pretenden ser una entrada de usuario. Por ejemplo, si el usuario está actuando de manera autoritaria y/o directiva (p.ej. pausado, serio, sin animación facial), se puede indicar que las palabras pretendían ser una entrada de usuario.

En 318, el método 300 comprende la determinación a partir de la comparación en 316 de si las palabras pronunciadas pretenden ser una entrada de usuario basándose en la información contextual. Si se determina que las palabras pretendían ser una entrada de voz, entonces el método 300 incluye, en 320, la realización a través del dispositivo informático de la acción asociada con la entrada de voz. De manera similar, si se determina que las palabras no pretendían ser una entrada de voz, entonces el método 300 incluye, en 322, no realizar una acción a través del dispositivo informático en respuesta a las palabras.

La FIG. 4 muestra un diagrama de flujo que representa una realización de un método 400 para la identificación de palabras pronunciadas procedentes de una combinación de audio e información de profundidad. El método 400 puede ser realizado por un dispositivo informático configurado para recibir audio y entrada de profundidad, tal como el sistema informático 102.

En 402, el método 400 incluye la recepción de información de profundidad procedente de una cámara de profundidad, y en 404, la recepción de información de audio procedente de uno o más micrófonos. En 406, uno o

más de los movimientos de la boca, la lengua, y la garganta del usuario se localizan desde la información de profundidad. Por ejemplo, la obtención del rasgo puede hacerse sobre la información de profundidad para determinar dónde está localizada cada rasgo facial listado previamente.

En 408, los movimientos de la boca, la lengua, y/o la garganta pueden ser identificados. Por ejemplo, un grado de apertura de la boca del usuario, la posición/forma de la lengua, la forma/localización de los labios del usuario, etc. cuando el usuario habla pueden ser rastreados para identificar los movimientos.

En 410, el método 400 incluye opcionalmente la activación de reconocimiento de voz para empezar de forma receptiva la detección de movimientos identificados de la boca, la lengua y/o la garganta que indican que el usuario está hablando. En este sentido, se puede evitar la operación de procesar un reconocimiento de voz con uso intensivo de recursos hasta que los movimientos identificados indican que el usuario está hablando realmente.

En 412, el método 400 incluye la identificación de una entrada de voz del usuario. Como se ha explicado previamente, la entrada de voz puede incluir un comando para que el dispositivo de ordenador realice una acción, o puede incluir una entrada que ha de ser visualizada (p.ej. como texto) en un dispositivo de visualización y/o ser guardada. La identificación de la entrada de voz puede incluir por ejemplo, la identificación de una o más posibles palabras pronunciadas desde la información de audio en 414. La entrada de voz puede ser identificada desde los datos de audio de cualquier manera adecuada. Además, como se indica en 416, la identificación de la entrada de voz puede incluir la identificación de uno o más sonidos, palabras, y/o fragmentos de palabras posibles desde la información de profundidad. Por ejemplo, los movimientos de la boca, la lengua, y/o la garganta del usuario pueden ser utilizados para identificar sonidos, palabras, etc.

La identificación de la entrada de voz puede también incluir, en 418, la comparación de una o más posibles palabras pronunciadas identificadas desde la información de audio con una o más posibles palabras o sonidos pronunciados identificados desde la información de profundidad. Esto puede ayudar a incrementar la fiabilidad de las posibles palabras pronunciadas identificadas a través de los datos de audio, para ayudar a eliminar la ambigüedad de posibles ambigüedades del discurso (por ejemplo, para identificar los límites entre palabras a través del análisis del movimiento de la mano), para identificar palabras adicionales que se han perdido en los datos de audio, y/o puede ser usado de cualquier otra manera adecuada.

Como un ejemplo más específico, los movimientos de la boca, la lengua, y/o la garganta del usuario pueden ser analizados (p.ej. extrayendo los datos de movimiento desde las imágenes de profundidad y aplicando una o más funciones de clasificación a los datos de movimiento) para identificar posibles palabras/sonidos pronunciados. Además, en algunas realizaciones, las puntuaciones de confianza pueden ser aplicadas a posibles palabras/sonidos pronunciados. Entonces, las determinadas posibles palabras/sonidos pronunciados determinadas desde la información de profundidad pueden ser comparados con las posibles palabras pronunciadas determinadas desde la información de audio, que de manera similar pueden incluir datos de la puntuación de confianza en algunas realizaciones. A partir de esta comparación, pueden identificarse con mayor probabilidad una palabra o palabras pronunciadas, p.ej. desde una puntuación de confianza combinada mayor, u otra métrica adecuada. Se entenderá que cualquier mecanismo adecuado puede ser utilizado para la comparación de posibles sonidos/palabras pronunciados identificados a través de la información de profundidad y las posibles palabras pronunciadas identificadas a través de la información de audio.

En 420, el método 400 incluye realizar una acción basada en la entrada de voz. Como se describió anteriormente, se puede realizar cualquier acción adecuada. Por ejemplo, la voz identificada puede ser utilizada como un comando de entrada para causar que el dispositivo de ordenador realice una acción, puede ser mostrada y/o guardada como contenido, puede ser utilizada para marcar contenido basado en un estado emocional determinado del usuario cuando está hablando, y/o cualquier otra acción adecuada.

En algunas realizaciones, los métodos y los procesos descritos anteriormente pueden estar ligados a un sistema informático que incluye uno o más ordenadores. En particular, los métodos y los procesos descritos aquí pueden ser implementados como una aplicación de ordenador, un servicio de ordenador, API de ordenador, librería de ordenador, y/u otro producto programa de ordenador.

La FIG. 5 muestra esquemáticamente una realización no limitativa de un sistema informático 500 que puede adoptar uno o más de los métodos y procesos descritos anteriormente. El sistema informático 500 es un ejemplo no limitativo de sistema informático 102. El sistema informático 500 es mostrado de forma simplificada. Se comprenderá que virtualmente cualquier arquitectura informática puede ser utilizada sin apartarse del alcance de esta descripción. En diferentes realizaciones, el sistema informático 500 puede tomar la forma de un ordenador central, ordenador servidor, ordenador de sobremesa, ordenador portátil, tableta, ordenador de entretenimiento para el hogar, dispositivo informático en red, dispositivo de juegos, dispositivo informático móvil, dispositivo de comunicación móvil (p.ej. teléfono inteligente), etc.

El sistema informático 500 incluye un subsistema 502 lógico y un subsistema 504 de almacenamiento. El sistema informático 500 puede opcionalmente incluir un subsistema 506 de visualización, subsistema 508 de entrada, subsistema 510 de comunicación, y/u otros componentes no mostrados en la FIG. 5.

El subsistema 502 lógico incluye uno o más dispositivos físicos configurados para ejecutar instrucciones. Por ejemplo, el subsistema lógico puede ser configurado para ejecutar instrucciones que son parte de una o más aplicaciones, servicios, programas, rutinas, librerías, objetos, componentes, datos estructurados, u otras construcciones lógicas. Tales instrucciones pueden ser implementadas para realizar una tarea, implementar un tipo de dato, transformar el estado de uno o más componentes, o llegar a un resultado deseado de otra manera.

El subsistema lógico puede incluir uno o más procesadores configurados para ejecutar instrucciones de software. Adicional o alternativamente, el subsistema lógico puede incluir una o más máquinas lógicas de hardware o firmware configuradas para ejecutar instrucciones de hardware o firmware. Los procesadores del subsistema lógico pueden ser de un núcleo o varios núcleos, y los programas ejecutados al respecto pueden ser configurados para procesamiento secuencial, paralelo o distribuido. El subsistema lógico puede opcionalmente incluir componentes individuales que están distribuidos entre dos o más dispositivos, que pueden estar ubicados y/o configurados remotamente para procesamiento coordinado. Aspectos del subsistema lógico pueden ser hechos virtuales y ejecutados por, dispositivos informáticos en red accesibles remotamente configurados en una configuración de computación en la nube.

El subsistema 504 de almacenamiento incluye uno o más dispositivos físicos, no transitorios configurados para contener datos y/o instrucciones ejecutables por el subsistema lógico para implementar los métodos y procesos descritos aquí. Cuando tales métodos y procesos son implementados, el estado del subsistema 504 de almacenamiento puede ser transformado – p.ej. para contener diferentes datos.

El subsistema 504 de almacenamiento puede incluir medios extraíbles y/o dispositivos integrados. El subsistema 504 de almacenamiento puede incluir dispositivos de memoria óptica (p.ej. CD, DVD, HD-DVD, disco Blu-Ray, etc.), dispositivos de memoria semiconductores (por ejemplo RAM, EPROM, EEPROM. etc.) y/o dispositivos de memoria magnéticos (p.ej. dispositivo de disco duro, dispositivo de disquete, dispositivo de cinta, MRAM, etc.), entre otros. El subsistema 504 de almacenamiento puede incluir dispositivos volátiles, no volátiles, dinámicos, estáticos, de lectura/escritura, de sólo lectura, de acceso aleatorio, de acceso secuencial, de localización direccionable, de fichero direccionable, y/o de contenido direccionable.

Se apreciará que el subsistema 504 de almacenamiento incluye uno o más dispositivos físicos, no transitorios. Sin embargo, en algunas realizaciones, aspectos de las instrucciones descritas aquí pueden propagarse en una forma transitoria por una pura señal (p.ej. una señal electromagnética, una señal óptica, etc.) que no está contenida por un dispositivo físico durante una duración finita. Además, los datos y/u otras formas de información que pertenecen a la presente descripción pueden ser propagados por una señal pura.

En algunas realizaciones, aspectos del subsistema 502 lógico y del subsistema 504 de almacenamiento pueden ser integrados juntos dentro de uno o más componentes lógicos de hardware a través de los cuales se puede adoptar la funcionalidad descrita aquí. Tales componentes lógicos de hardware pueden incluir agrupaciones de puertas programables en campo (FPGA), circuitos integrados específicos de programa y aplicación (PASIC/ASIC), productos estándares específicos de programa y aplicación (PSSP/ASSP), sistemas de sistema en chip (SOC), y dispositivos lógicos complejos programables (CPLD), por ejemplo.

El término "módulo" puede ser utilizado para describir un aspecto del sistema informático 500 implementado para realizar una función particular. En algunos casos, un módulo puede ser ejemplificado a través del subsistema 502 lógico ejecutando instrucciones contenidas por el subsistema 504 de almacenamiento. Se entenderá que módulos diferentes pueden ser ejemplificados a partir de la misma aplicación, servicio, bloque de código, objeto, librería, rutina, API, función, etc. De modo similar, el mismo módulo puede ser ejemplificado por diferentes aplicaciones, servicios, bloques de código, objetos, rutinas, API, funciones, etc. El término "módulo" puede abarcar ficheros ejecutables individuales o en grupos, ficheros de datos, librerías, controladores, secuencias de comandos, registros de base de datos, etc.

Se apreciará que un “servicio”, como se usa aquí, es un programa de aplicación ejecutable a través de múltiples sesiones de usuario. Un servicio puede estar disponible para uno o más componentes del sistema, programas, y/u otros servicios. En algunas implementaciones, un servicio puede operar en uno o más dispositivos de ordenador servidor.

Cuando se incluye, el subsistema 506 de visualización puede ser utilizado para mostrar una representación visual de datos contenidos en el subsistema 504 de almacenamiento. Esta representación visual puede tomar la forma de una interfaz gráfica de usuario (GUI). Como los métodos y procesos aquí descritos cambian los datos contenidos por el subsistema de almacenamiento, y así transforman el estado del subsistema de almacenamiento, el estado del subsistema 506 de visualización puede ser transformado de modo similar para representar visualmente cambios en los datos subyacentes. El subsistema 506 de visualización puede incluir uno o más dispositivos de visualización que utilizan virtualmente cualquier tipo de tecnología. Tales dispositivos de visualización pueden ser combinados con el subsistema 502 lógico y/o el subsistema 504 de almacenamiento en un alojamiento compartido, o tales dispositivos de visualización pueden ser dispositivos de visualización periféricos.

Cuando se incluye, el subsistema 508 de entrada puede comprender o interconectar con uno o más dispositivos de entrada de usuario tal como un teclado, un ratón, una pantalla táctil, o un controlador de juegos. En algunas realizaciones, el subsistema de entrada puede comprender o interconectar con componentes de entrada de usuario natural (NUI) seleccionados. Tales componentes pueden estar integrados o ser periféricos, y la transducción y/o procesamiento de acciones de entrada pueden ser manejados interna o externamente. Ejemplo de componentes NUI pueden incluir uno o más micrófonos para reconocimiento de discurso y/o de voz; una cámara infrarroja, de color, estereoscópica, y/o de profundidad para la visión automática y/o reconocimiento de gestos; un rastreador de cabeza, un rastreador de ojos, un acelerómetro, y/o giroscopio para detección de movimiento y/o reconocimiento de intención; así como componentes de detección de campo eléctrico para evaluación de la actividad cerebral.

Cuando se incluye, el subsistema 510 de comunicación puede estar configurado para acoplar de forma comunicativa el sistema informático 500 con uno o más dispositivos informáticos. El subsistema 510 de comunicación puede incluir dispositivos de comunicación cableados y/o inalámbricos compatibles con uno o más protocolos de comunicaciones diferentes. Como ejemplos no limitativos, el subsistema de comunicación puede estar configurado para la comunicación a través de una red telefónica inalámbrica, o una red local o de red de área amplia cableada o inalámbrica. En algunas realizaciones, el subsistema de comunicación puede permitir al sistema informático 500 enviar y/o recibir mensajes a y/o desde otros dispositivos a través de una red tal como la Internet.

Además, el sistema informático 500 puede incluir un módulo 512 de modelado estructural configurado para recibir información de imagen desde una cámara de profundidad 520 (descrita a continuación) e identificar y/o interpretar uno o más gestos realizados por un usuario. El sistema informático 500 puede también incluir un módulo 514 de reconocimiento de voz para identificar y/o interpretar uno o más comandos de voz o palabras pronunciadas emitidas por el usuario detectadas a través de uno o más micrófonos (acoplados al sistema informático 500 o a la cámara de profundidad). Mientras el módulo 512 de modelado estructural y el módulo 514 de reconocimiento de voz se representan como integrados dentro del sistema informático 500, en algunas realizaciones, uno o ambos de los módulos pueden sin embargo estar incluidos dentro de la cámara de profundidad 520.

El sistema informático 500 puede estar acoplado operativamente a la cámara de profundidad 520. La cámara de profundidad 520 puede incluir una luz infrarroja 522 y una cámara de profundidad 524 (también referida como una cámara de luz infrarroja) configurada para obtener vídeo de una escena que incluye uno o más sujetos humanos. El video puede comprender una secuencia de imágenes resuelta en el tiempo de resolución espacial y una frecuencia de fotogramas adecuada para el propósito aquí expuesto. Como se describe anteriormente con referencia a la FIG. 1, la cámara de profundidad y/o un sistema informático cooperante (p.ej. sistema informático 500) pueden ser configurados para procesar el vídeo obtenido para identificar una o más posturas y/o gestos del usuario, determinar una posición y un seguimiento de los movimientos de la boca, la lengua, y/o la garganta de un usuario, e interpretar tales posturas y/o gestos como comandos de dispositivo configurados para controlar varios aspectos del sistema informático 500.

La cámara de profundidad 520 puede incluir un módulo 526 de comunicación configurado para acoplar de forma comunicativa la cámara de profundidad 520 con uno o más dispositivos informáticos. El módulo 526 de comunicación puede incluir dispositivos de comunicación cableados y/o inalámbricos compatibles con uno o más protocolos diferentes de comunicaciones. En una realización, el módulo 526 de comunicación puede incluir una interfaz 528 de formación de imágenes para enviar información de imagen (tal como vídeo obtenido) al sistema informático 500. Adicionalmente o alternativamente, el módulo 526 de comunicación puede incluir una interfaz 530 de control para recibir instrucciones desde el sistema informático 500. Las interfaces de control e imagen pueden ser previstas como interfaces separadas, o pueden ser la misma interfaz. En un ejemplo, la interfaz 530 de control y la interfaz 528 de imagen pueden incluir un bus serie universal.

La naturaleza y número de cámaras puede diferir en varias cámaras de profundidad de acuerdo al alcance de esta descripción. En general, una o más cámaras pueden ser configuradas para proporcionar vídeo a partir del cual se obtiene una secuencia resuelta en el tiempo de mapas de profundidad tridimensionales a través de la transformación aguas abajo. Como se usa aquí, el término "mapa de profundidad" se refiere a una matriz de pixeles registrados en las regiones correspondientes de una escena visualizada, indicando un valor de profundidad de cada pixel la profundidad de la superficie visualizada por ese pixel. "Profundidad" se define como una coordenada paralela al eje óptico de la cámara de profundidad, que aumenta con el aumento de la distancia desde la cámara de profundidad.

En algunas realizaciones, la cámara de profundidad 520 puede incluir cámaras estereoscópicas derecha e izquierda. Las imágenes resueltas en el tiempo procedentes de ambas cámaras pueden ser hechas coincidir entre sí y ser combinadas para obtener video resuelto de profundidad.

En algunas realizaciones, una cámara de profundidad de "luz estructurada" puede ser configurada para proyectar una iluminación estructurada infrarroja que comprende numerosas, características discretas (p.ej. líneas o puntos). Una cámara puede ser configurada para visualizar la iluminación estructurada reflejada desde la escena. Basado en el espacio entre características adyacentes en las diferentes zonas de la escena visualizada, se puede construir un mapa de profundidad de la escena.

En algunas realizaciones, una cámara de profundidad de "tiempo de vuelo" puede incluir una fuente de luz configurada para proyectar una iluminación infrarroja pulsada sobre la escena. Dos cámaras pueden ser configuradas para detectar la iluminación pulsada reflejada desde la escena. Las cámaras pueden incluir un obturador electrónico sincronizado con la iluminación pulsada, pero los tiempos de integración para las cámaras pueden diferir, de tal manera que un pixel resuelto en el tiempo de vuelo de la iluminación pulsada, desde la fuente de luz hacia la escena y después hacia las cámaras, es discernible por las cantidades relativas de la luz recibida en los píxeles correspondientes de las dos cámaras.

La cámara de profundidad 520 puede incluir una cámara 532 de luz visible (p.ej. cámara RGB). Imágenes resueltas en el tiempo procedentes de cámaras de color y de profundidad pueden ser hechas coincidir entre sí y combinadas para producir video en color resuelto en profundidad. La cámara de profundidad 520 y/o el sistema informático 500 pueden además incluir uno o más micrófonos 534. Uno o más micrófonos pueden determinar sonidos direccionales y/o no direccionales provenientes de usuarios en el espacio físico y/u otras fuentes. Los datos de audio pueden ser grabados por uno o más micrófonos 534. Tales datos de audio pueden ser determinados de cualquier manera adecuada sin apartarse del alcance de esta descripción.

Mientras que la cámara de profundidad 520 y el sistema informático 500 se muestran en la FIG. 5 como dispositivos separados, en algunas realizaciones la cámara de profundidad 520 y el sistema informático 500 pueden estar incluidos en un único dispositivo. Así, la cámara de profundidad 520 puede opcionalmente incluir un sistema informático 500.

Se entenderá que las configuraciones y/o aproximaciones descritas aquí son de naturaleza ejemplar, y que estas realizaciones específicas o ejemplos no deben ser considerados en un sentido limitativo, porque las posibles variaciones son numerosas. Las rutinas o métodos específicos descritos aquí pueden representar uno o más de cualquier número de estrategias de procesamiento. Como tal, varios actos ilustrados y/o descritos pueden ser realizados en la secuencia ilustrada y/o descrita, en otras secuencias, en paralelo, u omitidos. Además, el orden de los procesos descritos anteriormente puede ser cambiado.

REIVINDICACIONES

1. Un método (200) para el reconocimiento de voz de un usuario en un dispositivo informático, que comprende:

recibir (202) información de profundidad de un espacio físico desde una cámara de profundidad (520);

identificar elementos contextuales en la información de profundidad;

recibir (204) información de audio desde uno o más micrófonos (534);

identificar (206) un conjunto de una o más posibles palabras pronunciadas desde la información de audio;

determinar (208) una entrada de voz para el dispositivo informático basado en la comparación del conjunto de una o más posibles palabras pronunciadas desde la información de audio y los elementos contextuales, en donde determinar (208) la entrada de voz incluye eliminar la ambigüedad, en el conjunto de una o más posibles palabras pronunciadas, una o más de

- un homónimo y

- una palabra compuesta para distinción de las palabras individuales que componen la palabra compuesta; y

- realizar (214) una acción en el dispositivo informático basada en la entrada de voz determinada.

2. El método (200) de reivindicación 1, en donde identificar elementos contextuales comprende además identificar elementos contextuales en una o más de la información de audio desde un micrófono direccional y de la información de imagen desde una cámara (532) de luz visible.

3. El método (200) de reivindicación 2, donde identificar los elementos contextuales comprende una o más de la determinación de una identidad de usuario basada en una o más de la información de profundidad e información desde una cámara (532) de luz visible, determinación de un estado emocional del usuario, determinación de un estado físico del usuario, determinación de un gesto realizado por el usuario, e identificación de uno o más objetos en un espacio físico del usuario.

4. El método (200) de reivindicación 1, que comprende además identificar un conjunto de uno o más posibles sonidos y/o palabras pronunciados desde la información de profundidad y comparar el conjunto de una o más posibles palabras pronunciadas identificadas a través de la información de audio con el conjunto de uno o más posibles sonidos y/o palabras pronunciados identificados a través de la información de profundidad para determinar la entrada de voz.

5. El método (200) de reivindicación 4, en donde identificar el conjunto de uno o más posibles sonidos y/o palabras pronunciadas desde la información de profundidad comprende además identificar uno o más movimientos de la boca, la lengua, y/o la garganta del usuario, e identificar el conjunto de uno o más posibles sonidos y/o palabras pronunciados basado en los movimientos.

6. El método (200) de reivindicación 1, en donde la entrada de voz comprende un comando y en donde realizar la acción comprende ejecutar el comando.

7. El método (200) de reivindicación 1, comprende además identificar qué usuario de una pluralidad de usuarios está hablando basado en uno o más movimientos de la boca y dirección de la mirada.

8. El método (200) de reivindicación 1, en donde la entrada de voz es contenido para ser almacenada y donde realizar la acción comprende almacenar el contenido.

9. El método (200) de reivindicación 1, en donde la entrada de voz comprende contenido que ha de ser visualizado en un dispositivo de visualización, y donde realizar la acción comprende enviar el contenido al dispositivo de visualización.

10. El método (200) de reivindicación 1, en donde un límite entre posibles sonidos y/o palabras pronunciados se determina en base a la identificación de los movimientos de la mano del usuario.

11. Un sistema informático (500) que comprende

un subsistema (502) lógico que incluye uno más dispositivos físicos configurados para ejecutar instrucciones; y

un subsistema (504) de almacenamiento que comprende uno o más dispositivos físicos, no transitorios, configurado para almacenar datos y/o instrucciones ejecutables por el subsistema (502) lógico para implementar el método (200) de cualquiera de las anteriores reivindicaciones.

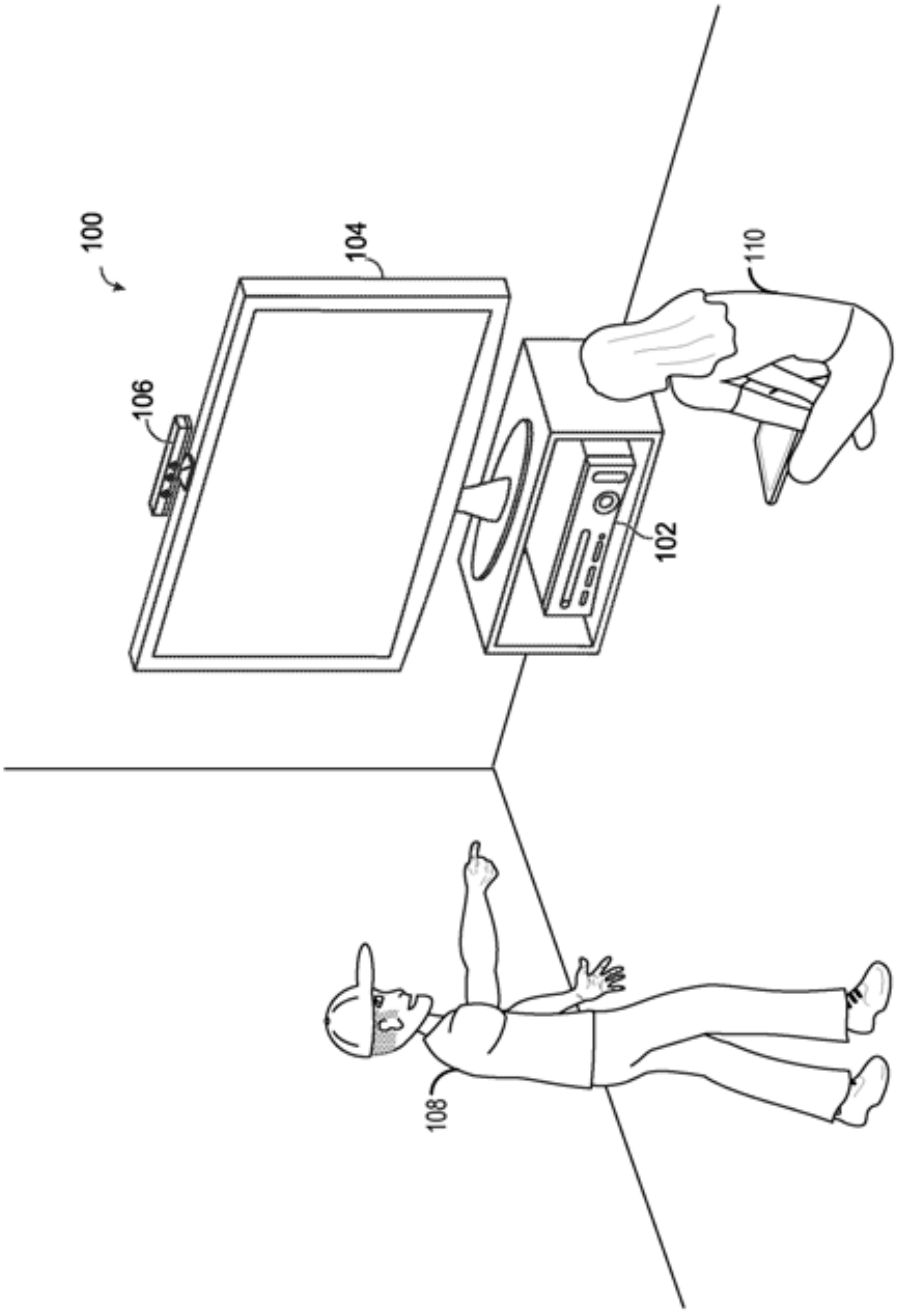

FIG. 1

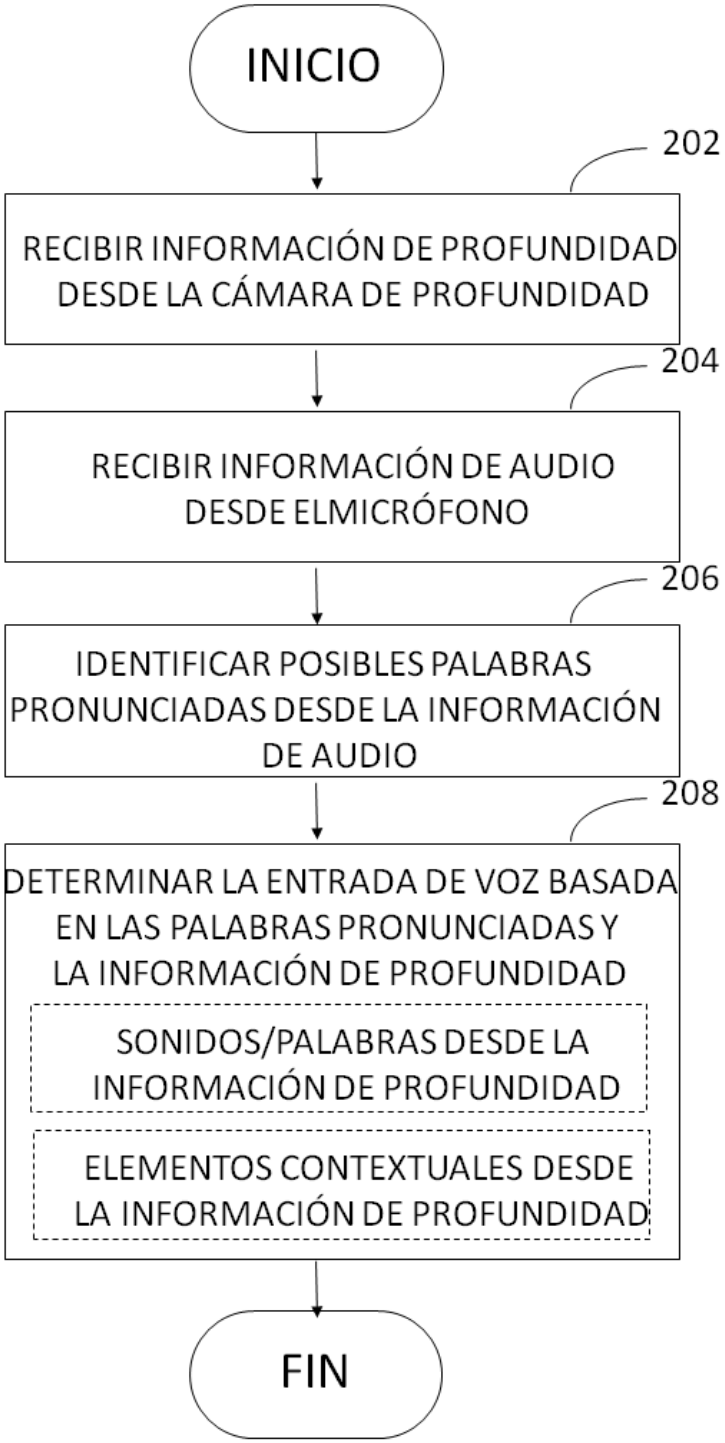
INICIO
202
RECIBIR INFORMACIÓN DE PROFUNDIDAD DESDE LA CÁMARA DE PROFUNDIDAD
204
RECIBIR INFORMACIÓN DE AUDIO DESDE ELMICRÓFONO
206
IDENTIFICAR POSIBLES PALABRAS PRONUNCIADAS DESDE LA INFORMACIÓN DE AUDIO
208
DETERMINAR LA ENTRADA DE VOZ BASADA EN LAS PALABRAS PRONUNCIADAS Y LA INFORMACIÓN DE PROFUNDIDAD
SONIDOS/PALABRAS DESDE LA INFORMACIÓN DE PROFUNDIDAD
ELEMENTOS CONTEXTUALES DESDE LA INFORMACIÓN DE PROFUNDIDAD
FIN

FIG. 2

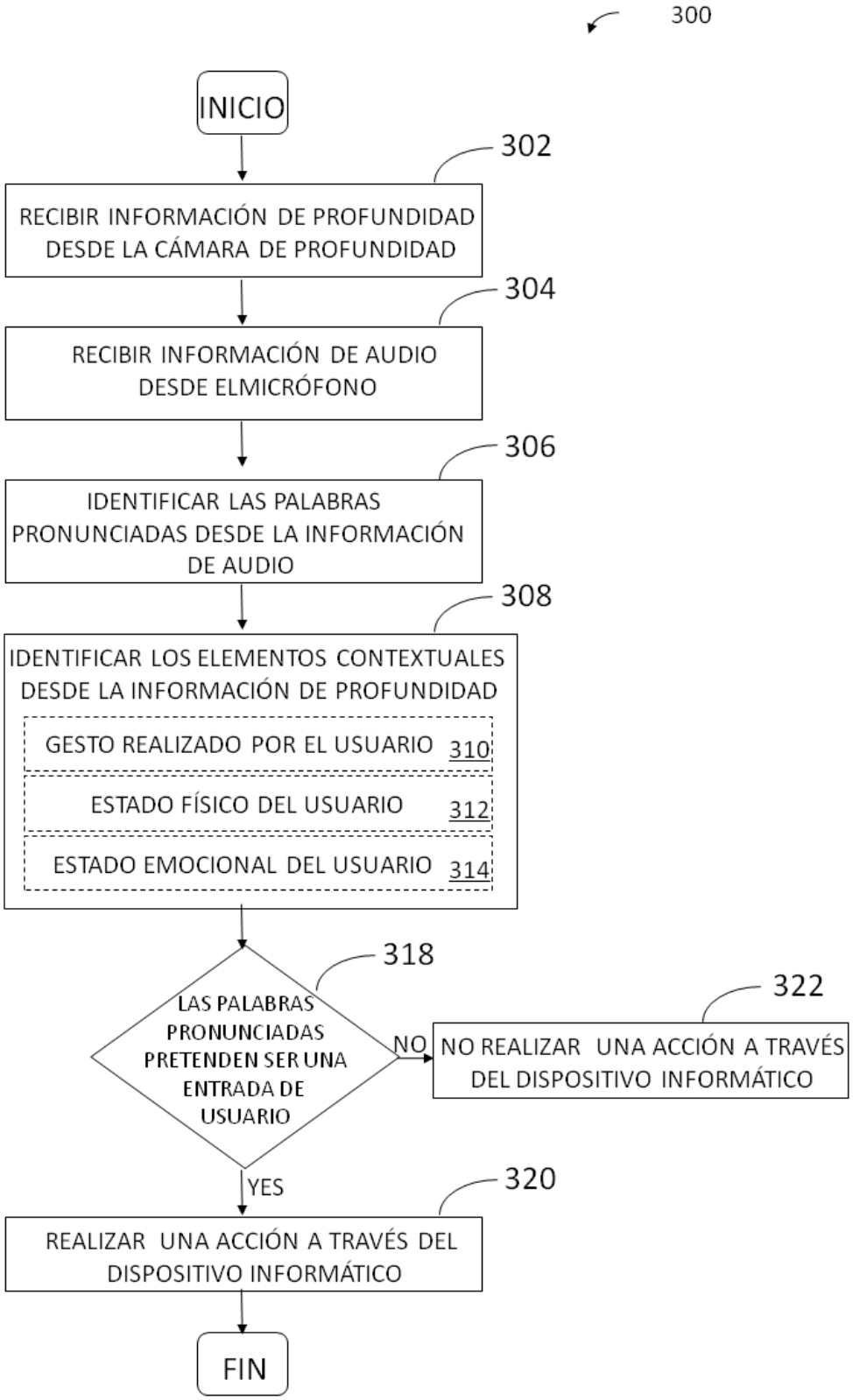
300
INICIO
302
RECIBIR INFORMACIÓN DE PROFUNDIDAD DESDE LA CÁMARA DE PROFUNDIDAD
304
RECIBIR INFORMACIÓN DE AUDIO DESDE ELMICRÓFONO
306
IDENTIFICAR LAS PALABRAS PRONUNCIADAS DESDE LA INFORMACIÓN DE AUDIO
308
IDENTIFICAR LOS ELEMENTOS CONTEXTUALES DESDE LA INFORMACIÓN DE PROFUNDIDAD
GESTO REALIZADO POR EL USUARIO 310
ESTADO FÍSICO DEL USUARIO 312
ESTADO EMOCIONAL DEL USUARIO 314
318
LAS PALABRAS PRONUNCIADAS PRETENDEN SER UNA ENTRADA DE USUARIO
NO
322
NO REALIZAR UNA ACCIÓN A TRAVÉS DEL DISPOSITIVO INFORMÁTICO
YES
320
REALIZAR UNA ACCIÓN A TRAVÉS DEL DISPOSITIVO INFORMÁTICO
FIN

FIG. 3

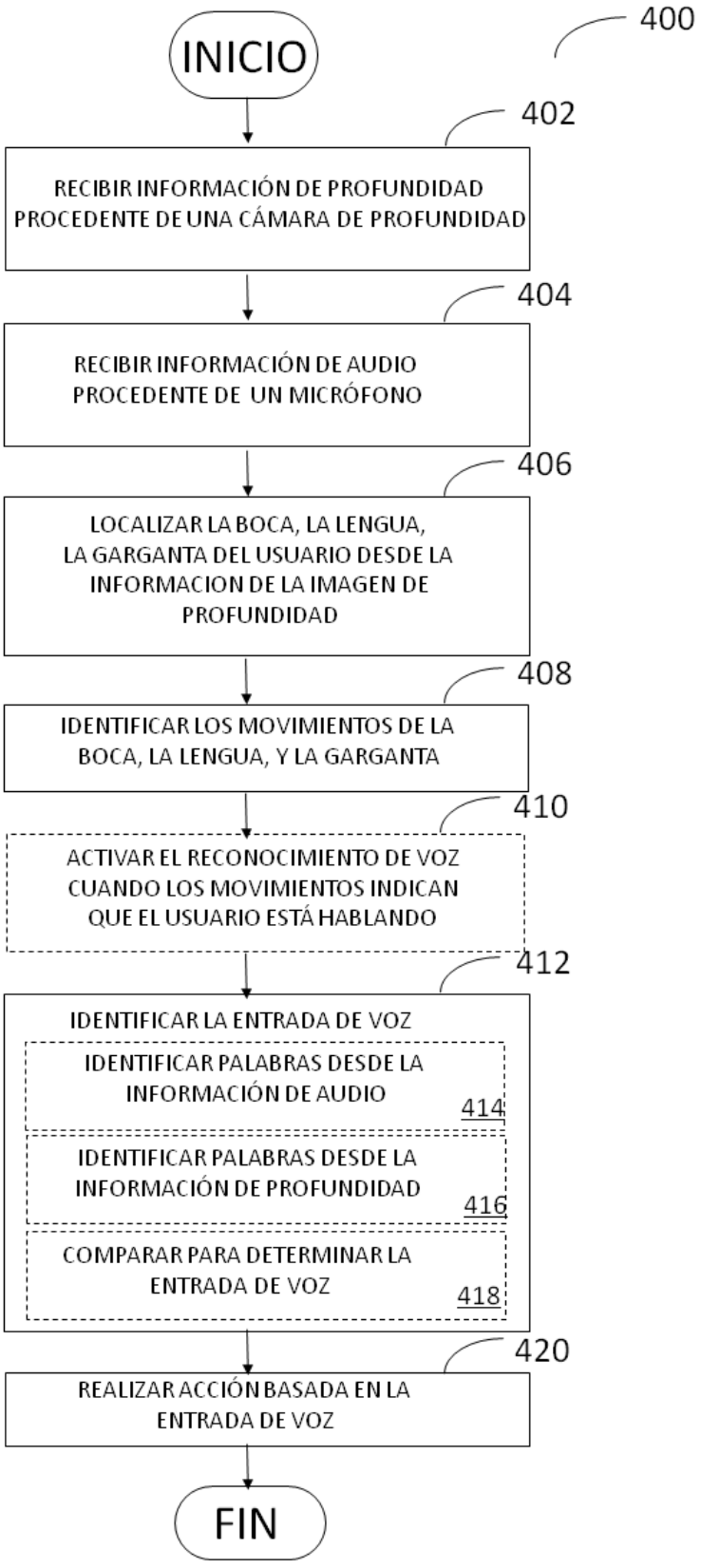
400
INICIO
402
RECIBIR INFORMACIÓN DE PROFUNDIDAD PROCEDENTE DE UNA CÁMARA DE PROFUNDIDAD
404
RECIBIR INFORMACIÓN DE AUDIO PROCEDENTE DE UN MICRÓFONO
406
LOCALIZAR LA BOCA, LA LENGUA, LA GARGANTA DEL USUARIO DESDE LA INFORMACION DE LA IMAGEN DE PROFUNDIDAD
408
IDENTIFICAR LOS MOVIMIENTOS DE LA BOCA, LA LENGUA, Y LA GARGANTA
410
ACTIVAR EL RECONOCIMIENTO DE VOZ CUANDO LOS MOVIMIENTOS INDICAN QUE EL USUARIO ESTÁ HABLANDO
412
IDENTIFICAR LA ENTRADA DE VOZ
IDENTIFICAR PALABRAS DESDE LA INFORMACIÓN DE AUDIO
414
IDENTIFICAR PALABRAS DESDE LA INFORMACIÓN DE PROFUNDIDAD
416
COMPARAR PARA DETERMINAR LA ENTRADA DE VOZ
418
420
REALIZAR ACCIÓN BASADA EN LA ENTRADA DE VOZ
FIN

FIG. 4

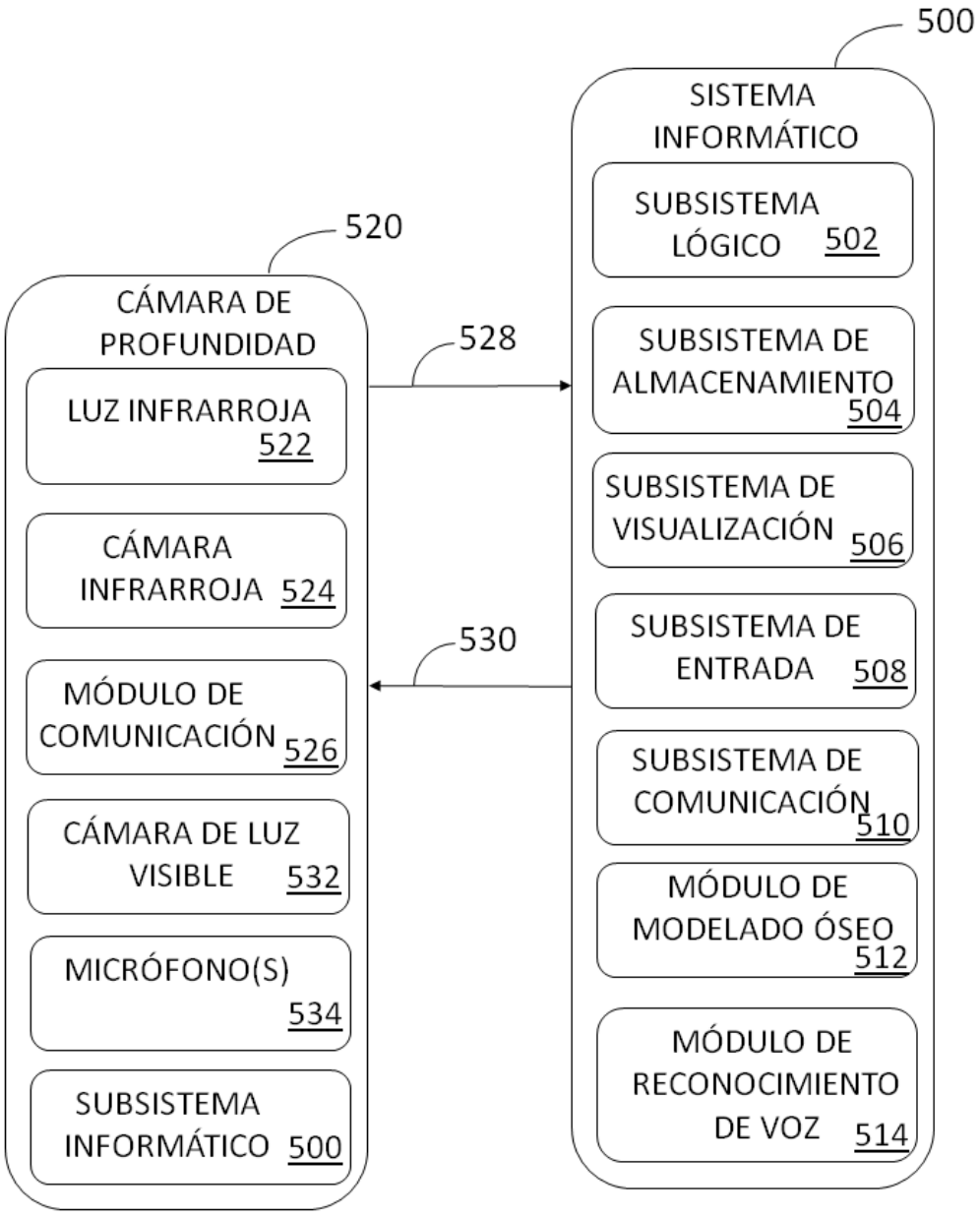
500
SISTEMA INFORMÁTICO
SUBSISTEMA LÓGICO 502
SUBSISTEMA DE ALMACENAMIENTO 504
SUBSISTEMA DE VISUALIZACIÓN 506
SUBSISTEMA DE ENTRADA 508
SUBSISTEMA DE COMUNICACIÓN 510
MÓDULO DE MODELADO ÓSEO 512
MÓDULO DE RECONOCIMIENTO DE VOZ 514
520
CÁMARA DE PROFUNDIDAD
LUZ INFRARROJA 522
CÁMARA INFRARROJA 524
MÓDULO DE COMUNICACIÓN 526
CÁMARA DE LUZ VISIBLE 532
MICRÓFONO(S) 534
SUBSISTEMA INFORMÁTICO 500
528
530

FIG. 5