(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 806 204**

(51) Int. CI.:

G10L 15/22 (2006.01)
G10L 15/08 (2006.01)
G10L 17/22 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **15.06.2016 PCT/US2016/037495**

(87) Fecha y número de publicación internacional: **21.12.2017 WO17217978**

(96) Fecha de presentación y número de la solicitud europea: **15.06.2016 E 16733794 (8)**

(97) Fecha y número de publicación de la concesión europea: **20.05.2020 EP 3472831**

(54) Título: **Técnicas para reconomiento de voz para activación y sistemas y métodos relacionados**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
16.02.2021

(73) Titular/es:

CERENCE OPERATING COMPANY (100.0%)
15 Wayside Road
Burlington, MA 01803, US

(72) Inventor/es:

PFEFFINGER, MEIK;
MATHEJA, TIMO;
HERBIG, TOBIAS y
HAULICK, TIM

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Técnicas para reconocimiento de voz para activación y sistemas y métodos relacionados

Antecedentes

Un sistema puede estar habilitado para hablar, lo que permite a los usuarios interaccionar con el sistema a través del habla, por ejemplo, al permitir que los usuarios digan comandos para controlar el sistema. Emplear un sistema controlado por voz a menudo requiere que los usuarios indiquen al sistema controlado por voz que el usuario tiene la intención de interaccionar con el sistema al hablar. Por ejemplo, un sistema controlado por voz puede configurarse para que comience a reconocer el habla después de un disparador manual, tal como presionar un botón (por ejemplo, un botón de un dispositivo físico y/o un botón dentro de una interfaz de usuario de una aplicación de software de reconocimiento de voz), el lanzamiento de una aplicación u otra interacción manual con el sistema se proporciona para alertar al sistema de que el discurso después del disparador manual está dirigido al sistema. Sin embargo, los disparadores manuales complican la interacción con el sistema controlado por voz y, en algunos casos, pueden ser incómodos o imposibles de usar (por ejemplo, cuando las manos del usuario están ocupadas de otra manera, como cuando se está conduciendo un vehículo o cuando el usuario está demasiado lejos del sistema como para conectarse manualmente con el sistema o una interfaz del mismo).

Para abordar los inconvenientes (y a menudo la falta de aplicabilidad) de los disparadores manuales, algunos sistemas controlados por voz permiten que se activen los disparadores de voz para comenzar a interactuar con el sistema, eliminando así al menos algunas (si no todas) acciones manuales y facilitando generalmente el acceso manos libres al sistema controlado por voz. Un disparador de voz puede comprender una palabra o frase designada (por ejemplo, "Hola Dragon") que el usuario pronuncia para indicarle al sistema controlado por voz que el usuario tiene la intención de interactuar con el sistema (por ejemplo, emitir uno o más comandos al sistema). Un disparador de voz también se conoce en el presente documento como "palabra de activación" o "WuW". Después de que se ha detectado una palabra de activación, el sistema comienza a reconocer el discurso posterior pronunciado por el usuario. En la mayoría de los casos, a menos y hasta que el sistema detecte la palabra de activación, el sistema asumirá que la entrada acústica recibida del entorno no está dirigida o destinada al sistema y no procesará la entrada acústica más allá. Como tal, es importante que un sistema controlado por voz pueda detectar, con un grado relativamente alto de precisión, cuando se ha pronunciado una palabra de activación. El documento CN 104 575 504 A desvela un sistema donde una señal acústica es procesada en paralelo por un reconocedor de voz y una unidad de identificación de huella de voz para determinar si se ha recibido un comando de activación de un usuario registrado.

Sumario

De acuerdo con la invención, se proporciona un sistema según se establece en la reivindicación 1, un método según se establece en la reivindicación 14 y un medio según se establece en la reivindicación 15. Las realizaciones preferentes se exponen en las reivindicaciones dependientes. Algunas realizaciones se refieren a un sistema para detectar al menos una palabra de activación designada para al menos una aplicación controlada por voz. El sistema comprende al menos un micrófono; y al menos un procesador de hardware de ordenador configurado para realizar: recibir una señal acústica generada por el al menos un micrófono, al menos en parte como resultado de recibir un enunciado pronunciado por una persona que habla; obtener información indicativa de la identidad de la persona que habla; interpretar la señal acústica al menos en parte determinando, utilizando la información indicativa de la identidad de la persona que habla y el reconocimiento automático de voz, si el enunciado pronunciado por la persona que habla incluye la al menos una palabra de activación designada; e interactuar con la persona que habla en función, al menos en parte, de los resultados de la interpretación.

Algunas realizaciones se refieren a un método para detectar al menos una palabra de activación designada para al menos una aplicación controlada por voz. El método comprende el uso de al menos un procesador de hardware de ordenador para realizar: recibir una señal acústica generada por al menos un micrófono al menos en parte como resultado de recibir un enunciado pronunciado por una persona que habla; obtener información indicativa de la identidad de la persona que habla; interpretar la señal acústica al menos en parte determinando, utilizando la información indicativa de la identidad de la persona que habla y el reconocimiento automático de voz, si el enunciado pronunciado por la persona que habla incluye la al menos una palabra de activación designada; e interactuar con la persona que habla en función, al menos en parte, de los resultados de la interpretación.

Algunas realizaciones se refieren al menos a un medio de almacenamiento legible por ordenador no transitorio que almacena instrucciones ejecutables por el procesador que, cuando son ejecutadas por al menos un procesador de hardware de ordenador, hacen que al menos un procesador de hardware de ordenador realice un método para detectar al menos una palabra de activación designada para al menos una aplicación controlada por voz. El método comprende: recibir una señal acústica generada por al menos un micrófono, al menos en parte como resultado de recibir un enunciado pronunciado por una persona que habla; obtener información indicativa de la identidad de la persona que habla; interpretar la señal acústica al menos en parte determinando, utilizando la información indicativa de la identidad de la persona que habla y el reconocimiento automático de voz, si el enunciado pronunciado por la

persona que habla incluye la al menos una palabra de activación designada; e interactuar con la persona que habla en función, al menos en parte, de los resultados de la interpretación.

Breve descripción de los dibujos

Se describirán diversos aspectos y realizaciones con referencia a las siguientes figuras. Las figuras no están necesariamente dibujadas a escala.

La figura 1 es un diagrama de bloques de un sistema ilustrativo controlado por voz, de acuerdo con algunas realizaciones de la tecnología descrita aquí.
La figura 2 es un diagrama de bloques de otro sistema ilustrativo controlado por voz, de acuerdo con algunas realizaciones de la tecnología descrita en el presente documento.
La figura 3 es un diagrama de flujo de un proceso ilustrativo para detectar una palabra de activación en un enunciado basado, al menos en parte, en información indicativa de la identidad de la persona que pronuncia el enunciado, de acuerdo con algunas realizaciones de la tecnología descrita en el presente documento.
La figura 4 es una ilustración de los datos asociados con una o más personas que hablan que pueden usarse para realizar la detección de la palabra de activación, de acuerdo con algunas realizaciones de la tecnología descrita en el presente documento.
La figura 5 es un diagrama de bloques de un sistema informático ilustrativo que puede usarse para implementar algunas realizaciones de la tecnología descrita en el presente documento.

Descripción detallada

Muchos sistemas controlados por voz permiten una participación generalmente de manos libres mediante el uso de palabras de activación. Una palabra de activación puede ser un enunciado hablado de una palabra, un enunciado hablado de varias palabras y/o cualquier enunciado hablado (de cualquier longitud adecuada que, por ejemplo, puede ser más corto que una sola palabra) que puede ser pronunciado por un usuario para indicar su intención de interactuar con un sistema controlado por voz. Dado que una palabra de activación debe reconocerse en general antes de que el sistema controlado por voz responda al usuario (por ejemplo, antes de que el sistema responda a otros comandos de voz), es deseable que la palabra de activación se reconozca con un alto grado de precisión. Las tasas de falsos positivos y falsos negativos que son demasiado altas dan como resultado un sistema con una capacidad de respuesta insatisfactoria, lo que conduce a la frustración y enfado del usuario. Como tal, los sistemas controlados por voz se benefician de la detección robusta de palabras de activación.

A medida que los entornos controlados por voz se vuelven cada vez más sofisticados, es posible que los sistemas controlados por voz necesiten responder a múltiples oradores diferentes que pueden tratar de participar y/o interactuar con múltiples aplicaciones. Por ejemplo, un vehículo (por ejemplo, un automóvil) puede incluir un sistema de telefonía manos libres, un sistema de navegación del vehículo, un sistema de sonido, un sistema de televisión y/o uno o más componentes controlables del vehículo (por ejemplo, ventanas, control de clima, etc.) que el conductor y/o los pasajeros pueden desear controlar mediante el habla. Como otro ejemplo, una casa inteligente o una habitación inteligente puede incluir un televisor, sistema de sonido, sistema de iluminación, control de clima, sistema de seguridad y/u otros sistemas con los que uno o más ocupantes pueden tratar de interactuar a través del habla. Muchos sistemas convencionales están configurados para detectar una sola palabra de activación y pueden ser capaces de hacerlo satisfactoriamente en entornos en los que solo está hablando un solo usuario. Dichos sistemas convencionales pueden ser inadecuados para entornos que tienen múltiples oradores, que pueden hablar simultáneamente o en proximidad cercana y/o que potencialmente buscan involucrarse en diferentes aspectos del sistema (por ejemplo, diferentes aplicaciones o sistemas controlados por voz en el mismo entorno).

Los inventores han reconocido que la capacidad de diferenciar entre las personas que hablan no solo facilita una detección de palabras de activación más sólida, sino que también puede proporcionar a cada persona que habla una interacción más personalizada con un sistema controlado por voz, por ejemplo, a través de la personalización de las palabras de activación y/u otra personalización del sistema controlado por voz a la persona que habla. En consecuencia, en algunas realizaciones, un sistema controlado por voz puede configurarse para obtener información indicativa de la identidad de una persona que habla y usar la información obtenida para mejorar la detección de palabras de activación y hacerlo más sólido, para mejorar la calidad de la interacción entre la persona que habla y el sistema, y / o para cualquier otro propósito o propósitos adecuados ejemplos de los cuales se proporcionan a continuación.

La información indicativa de la identidad de una persona que habla puede incluir cualquier información que pueda usarse para determinar la identidad de la persona que habla y/o para diferenciar entre la persona que habla y una o más personas, y en algunas realizaciones, un sistema controlado por voz puede usar la información obtenida indicativa de la identidad de una persona que habla para hacerlo. En algunas realizaciones, la información indicativa de la identidad de la persona que habla puede incluir información relacionada con las características del discurso de la persona que habla que podría usarse para identificar la identidad de la persona que habla. Por ejemplo, un sistema controlado por voz puede comparar las características del discurso almacenadas (por ejemplo, una "impresión de voz" almacenada) de una persona que habla, que es conocido por el sistema (por ejemplo, al estar

registrado en el sistema), con las características del discurso obtenidas de la entrada acústica recibida del entorno para determinar si la entrada acústica incluye voz de la persona que habla. En algunas realizaciones, cuando las características de una persona que habla no coinciden con ninguna característica de voz almacenada por el sistema, el sistema puede permitir que la persona que habla se registre en el sistema. De forma adicional o alternativa, la información indicativa de la identidad de la persona que habla puede incluir cualquier información sobre el comportamiento de la persona que habla que pueda usarse para determinar la identidad de la persona que habla. Un ejemplo de dicha información sobre el comportamiento de la persona que habla es la información que indica dónde se encuentra generalmente una persona que habla cuando se emiten comandos a un sistema controlado por voz. Por ejemplo, una persona que habla dada normalmente puede ser el conductor de un vehículo en particular y un sistema controlado por voz en el vehículo puede determinar la identidad de esta persona que habla (o la identidad probable de la persona que habla) al determinar que la entrada de voz se recibió desde el asiento del conductor. Del mismo modo, la ubicación habitual de una persona en una habitación inteligente se puede utilizar para determinar la identidad de una persona que habla. También se puede utilizar otra información indicativa de la identidad de la persona que habla, ya que los aspectos de la tecnología descritos en el presente documento no están limitados a este respecto. Por ejemplo, en algunas realizaciones, el sistema puede aprender los hábitos de una o más personas que hablan a lo largo del tiempo y puede usar información sobre los hábitos de una persona que habla para identificar a una persona que habla y/o mejorar la robustez de detectar las palabras de activación pronunciadas por la persona que habla.

Algunas realizaciones descritas en el presente documento abordan todos los problemas descritos anteriormente que los inventores han reconocido con los sistemas convencionales para realizar la detección de palabras de activación. Sin embargo, no todas las realizaciones descritas a continuación abordan cada uno de estos problemas y algunas realizaciones pueden no abordar ninguno de ellos. Como tal, debe apreciarse que las realizaciones de la tecnología descrita en el presente documento no se limitan a abordar todos o cualquiera de los problemas tratados anteriormente de los sistemas convencionales para realizar la detección de palabras de activación.

En algunas realizaciones, un sistema controlado por voz puede configurarse para detectar una palabra de activación particular al: recibir una señal acústica generada por uno o más micrófonos, al menos en parte como resultado de recibir un enunciado pronunciado por una persona que habla; obtener información indicativa de la identidad de la persona que habla; interpretar la señal acústica utilizando la información indicativa de la identidad de la persona que habla para determinar si el enunciado pronunciado por la persona que habla incluye la palabra de activación particular; e interactuar con la persona que habla en función, al menos en parte, de los resultados de la interpretación de la señal acústica. Por ejemplo, cuando el sistema controlado por voz determina que el enunciado pronunciado por la persona que habla incluye la palabra de activación particular, el sistema controlado por voz puede interactuar con la persona que habla permitiendo que la persona que habla controle (por ejemplo, pronunciando uno o más comandos de voz) una o más aplicaciones controladas por voz que se ejecutan en el sistema controlado por voz. Por otro lado, cuando el sistema controlado por voz determina que el enunciado pronunciado por la persona que habla no incluye una palabra de activación, el sistema controlado por voz puede no permitir que la persona que habla controle ninguna aplicación controlada por voz que se ejecute en el sistema controlado por voz.

En algunas realizaciones, se puede configurar un sistema controlado por voz para determinar la identidad de la persona que habla (por ejemplo, procesando el habla de la persona que habla, analizando el comportamiento de la persona que habla, comparando el comportamiento de la persona que habla con la información almacenada que caracteriza los hábitos de las personas que hablan, en función de, al menos en parte, la posición de la persona que habla en el entorno, y/o de cualquier otra manera adecuada) y use la identidad de la persona que habla para determinar si el enunciado pronunciado por la persona que habla incluye una palabra de activación particular. El sistema controlado por voz puede usar la identidad de la persona que habla para determinar si el enunciado pronunciado por la persona que habla contiene la palabra de activación particular de alguna manera adecuada. En algunas realizaciones, por ejemplo, el sistema controlado por voz puede almacenar una lista personalizada de una o más palabras de activación para cada una o múltiples de las personas que hablan. Después de que el sistema controlado por voz determina la identidad de una persona que habla, el sistema puede acceder a una lista de una o más palabras de activación asociadas con la persona que habla identificada y proporcionar el contenido de la lista a un reconocedor de voz para usar en la realización del reconocimiento automático de voz en el enunciado de la persona que habla para determinar si el enunciado contiene alguna palabra o palabras de activación en la lista de palabras de activación accedidas.

Proporcionar un reconocedor de voz con acceso a una lista "personalizada" de palabras de activación asociadas con una persona que habla puede mejorar significativamente la capacidad del reconocedor de voz para determinar con precisión si una persona que habla pronunció alguna de las palabras de activación de la lista. Por ejemplo, un reconocedor de voz puede usar la lista de palabras de activación para restringir su espacio de búsqueda a las palabras de la lista (en lugar de buscar entre todas las palabras posibles en el vocabulario del reconocedor) para hacer que su rendimiento de reconocimiento de palabras de activación sea más robusto, por ejemplo, al aumentar la probabilidad de que se reconozca una palabra de activación en la lista cuando se pronuncia, reduciendo la probabilidad de que se reconozca una palabra de activación en la lista cuando no se pronuncia (es decir, reducir la probabilidad de un falso positivo) y/o reducir la probabilidad de que una palabra de activación en la lista no se

reconozca cuando se pronuncia (es decir, reducir la probabilidad de que ocurra un falso negativo). De forma adicional o alternativa, el reconocedor de voz puede usar la lista personalizada de palabras de activación para adaptar uno o más de sus componentes (por ejemplo, uno o más modelos acústicos, uno o más modelos de pronunciación, uno o más modelos de idiomas, uno o más transductores de estado finitos, uno o más léxicos o vocabularios, y/o cualquier otro componente o componentes adecuados) para hacer que su rendimiento de reconocimiento de palabras de activación sea más robusto.

Como otro ejemplo, en algunas realizaciones, se puede usar una lista de palabras de activación asociadas con una persona que habla para compensar las interferencias (por ejemplo, ruido, habla de uno o más de otros oradores) presente en la señal acústica que contiene el enunciado de la persona que habla y/o por cualquier artefacto causado por dicha interferencia. Por ejemplo, cuando el sistema controlado por voz detecta solo una parte de una palabra de activación particular debido a la presencia de interferencia, la parte detectada se puede comparar con una o más entradas en la lista de palabras de activación para identificar que la palabra de activación en particular probablemente fue pronunciada por la persona que habla a pesar de que no se detectó toda la palabra de activación debido a la interferencia. Como ejemplo específico, una persona que habla puede pronunciar la frase "Hola, mi coche favorito" como una palabra de activación para un sistema de navegación con voz en el coche de la persona que habla. Sin embargo, debido al ruido acústico, es posible que el sistema de navegación controlado por voz solo haya detectado la parte "Hola, mi co" del enunciado debido al ruido. Comparando la parte detectada "Hola, mi co" con la entrada, "Hola, mi coche favorito" en la lista de palabras de activación asociadas con la persona que habla, el sistema de navegación controlado por voz puede determinar que, debido a que la parte detectada coincide con al menos una parte umbral de una palabra de detección en la lista, es probable que la persona que habla pronuncie la palabra de activación" "Hola, mu coche favorito" y puede permitir que la persona que habla controle verbalmente el sistema de navegación por voz. De esta manera, la palabra de activación de la persona que habla se reconoce a pesar de la presencia del ruido, y la experiencia de la persona que habla al interactuar con el sistema de navegación controlado por voz puede mejorarse.

Sin embargo, debe apreciarse que un sistema controlado por voz no se limita al uso de la identidad de la persona que habla únicamente para mejorar su rendimiento de reconocimiento de palabras de activación. Más bien, un sistema controlado por voz puede usar la identidad de una persona que habla (o, más generalmente, información indicativa de la identidad de la persona que habla) para otros fines además de o en lugar de mejorar la solidez de su rendimiento de reconocimiento de palabras de activación. En algunas realizaciones, por ejemplo, un sistema controlado por voz puede usar información indicativa de la identidad de una persona que habla para personalizar su interacción con la persona que habla. Por ejemplo, el sistema controlado por voz puede permitir que las personas que hablan tengan palabras de activación personalizadas. Por ejemplo, el sistema controlado por voz puede configurarse para permitir que una persona que habla particular use una palabra de activación designada específica para la persona que habla particular para despertar una aplicación controlada por voz. El sistema controlado por voz puede configurarse además para no permitir que ninguna persona que habla que no sea la persona que habla en particular use la palabra de activación designada para activar la aplicación controlada por voz. Para implementar dicha funcionalidad, en algunas realizaciones, un sistema controlado por voz puede determinar la identidad de una persona que habla y usar la identidad de la persona que habla para limitar las palabras de activación que el sistema reconocerá a las palabras de activación asociadas con la identidad de la persona que habla, como se describe en el presente documento. De esta manera, cada individuo puede seleccionar una palabra de activación diferente para participar en aspectos de un sistema controlado por voz (por ejemplo, un sistema de navegación o sonido en un vehículo, un sistema de televisión o iluminación en un hogar, etc.).

En algunas realizaciones, un sistema controlado por voz puede personalizar su interacción con una persona que habla deduciendo, basándose al menos en parte en información indicativa de la identidad de la persona que habla, una o más acciones a tomar cuando interactúa con la persona que habla. Por ejemplo, el sistema controlado por voz puede identificar, en función de la identidad de la persona que habla, información indicativa de las preferencias y/o intereses de la persona que habla y tomar una o más acciones basadas en la información indicativa de las preferencias y/o intereses de la persona que habla cuando interactúa con la persona que habla. Como ejemplo específico, cuando una perrona que habla dice: "reproducir mi música favorita", el sistema controlado por voz puede determinar la música favorita de la persona que habla en función de la identidad de la persona que habla y comienza la reproducción de esa música.

En algunas realizaciones, un sistema controlado por voz puede usar información indicativa de la identidad de una persona que habla para implementar medidas de control de acceso. Un sistema controlado por voz puede configurarse para ejecutar una o más aplicaciones controladas por voz e interactuar con una o varias personas que hablan diferentes. Sin embargo, se puede configurar un sistema controlado por voz para evitar que algunas personas que hablan controlen algunas aplicaciones controladas por voz. Un sistema controlado por voz puede configurarse para permitir que una persona que habla en particular no controle ninguna, algunas (por ejemplo, una o más pero no todas) o todas las aplicaciones controladas por voz que se ejecutan en el sistema. En consecuencia, en algunas realizaciones, un sistema controlado por voz puede determinar, usando información indicativa de la identidad de la persona que habla, si la persona que habla está autorizada para controlar una aplicación controlada por voz en particular y, cuando se determina que la persona que habla está autorizada para controlar la aplicación controlada por voz e particular, puede permitir que la persona que habla lo haga. Por otro lado, cuando el sistema

controlado por voz determina que una persona que habla no está autorizada para controlar la aplicación controlada por voz en particular, el sistema controlado por voz puede evitar que la persona que habla lo haga.

Por ejemplo, la información indicativa de la identidad de una persona que habla puede incluir la posición de la persona que habla dentro del entorno y el sistema controlado por voz puede determinar si la persona que habla está autorizada para controlar una aplicación controlada por voz en particular en función de, al menos en parte, la posición de la persona que habla. Como ejemplo específico, un sistema controlado por voz puede permitir que cualquier persona que habla sentada en el asiento del conductor de un automóvil controle una aplicación de navegación controlada por voz, pero no permite que ninguna persona que habla sentada en el asiento trasero de un automóvil controle la aplicación de navegación. Como otro ejemplo, la información indicativa de la identidad de una persona que habla puede incluir la identidad de la persona que habla y el sistema controlado por voz puede determinar si una persona que habla está autorizada para controlar una aplicación controlada por voz en particular en función de su identidad. Para este fin, el sistema controlado por voz puede mantener información que indique qué personas que hablan están (y/o no) autorizadas para controlar varias aplicaciones controladas por voz y usar esta información junto con la identidad de una persona que habla para determinar si la persona que habla está autorizada para controlar un aplicación controlada por voz en particular.

En algunas realizaciones, un sistema controlado por voz puede usar información indicativa de la identidad de una o más personas que hablan para procesar la entrada de voz emitida simultáneamente por varias personas que hablan diferentes. Dos personas que hablan pueden hablar simultáneamente cuando los períodos durante los cuales las personas que hablan están hablando, al menos parcialmente, se superponen. Por ejemplo, un sistema controlado por voz puede usar información sobre la posición y/o identidad de múltiples personas que hablan diferentes para procesar las entradas de voz proporcionadas por las diversas personas que hablan simultáneamente o muy cerca una de la otra. Como ejemplo específico, el conductor de un automóvil puede pronunciar una primera palabra de activación para una aplicación de navegación controlada por voz (por ejemplo, para obtener indicaciones para llegar a un destino) simultáneamente con un pasajero en el asiento trasero de un automóvil que pronuncia una segunda palabra de activación para una aplicación de telefonía controlada por voz (por ejemplo, para hacer una llamada telefónica). El sistema controlado por voz puede configurarse para procesar los enunciados del conductor y el pasajero, utilizando información que indique su posición y/o identidad, para determinar: (1) si el conductor pronunció una palabra de activación para la aplicación de navegación (o cualquier otra) (por ejemplo, accediendo a una lista personalizada de palabras de activación asociadas con el conductor); (2) si el conductor está autorizado para controlar la aplicación de navegación (por ejemplo, según la posición del conductor en el automóvil y/o la identidad del conductor); (3) si el pasajero pronunció una palabra de activación para la aplicación de telefonía (o cualquier otra) (por ejemplo, accediendo a una lista personalizada de palabras de activación asociadas con el pasajero); y/o (4) si los pasajeros están autorizados a controlar la aplicación de telefonía (por ejemplo, en función de la identidad y/o posición del pasajero en el automóvil).

A continuación se presentan descripciones más detalladas de varios conceptos relacionados con las técnicas de detección de palabras de activación y realizaciones de las mismas. Debe tenerse en cuenta que varios aspectos descritos en el presente documento pueden implementarse de cualquiera de numerosas maneras. Los ejemplos de implementaciones específicas se proporcionan en el presente documento solo con fines ilustrativos. Además, los diversos aspectos descritos en las realizaciones siguientes pueden usarse solos o en cualquier combinación, y no se limitan a las combinaciones descritas explícitamente en el presente documento.

La figura 1 es un diagrama de bloques de un sistema 100 controlado por voz ilustrativo, de acuerdo con algunas realizaciones de la tecnología descrita en el presente documento. El sistema 100 incluye micrófono(s) 112, sensor(es) 114 y aplicación(es) 116 controladas por voz, que pueden ser parte del entorno 110. El entorno 110 puede ser cualquier entorno adecuado donde un usuario pueda controlar en o más aplicaciones 116 controladas por voz. Por ejemplo, el entorno 110 puede ser un vehículo (por ejemplo, un automóvil, un autobús, una barca, etc.), una casa inteligente, una habitación inteligente o cualquier otro entorno adecuado. El entorno 110 puede incluir una o varias personas que hablan. El hecho de que la o las aplicaciones 116 controladas por voz formen parte del entorno 110 no requiere que estas aplicaciones se ejecuten en un procesador ubicado físicamente en el entorno 100. Por el contrario, una persona que habla en el entorno 110 solo necesita poder interactuar con una interfaz (por ejemplo, una interfaz controlada por voz) de una aplicación controlada por voz para que esa aplicación se considere como en el entorno 110, como se muestra en la figura 1.

En algunas realizaciones, el micrófono(s) 112 pueden incluir cualquier número y tipo de cualquier transductor(es) adecuados configurados para convertir ondas acústicas en señales eléctricas. De acuerdo con algunas realizaciones, el o los micrófonos 112 pueden incluir uno o más micrófonos de presión de sonido, micrófonos electret, micrófonos binaurales, micrófonos MEMS o combinaciones de los mismos. Sin embargo, debe apreciarse que se puede usar cualquier tipo de micrófono en cualquier combinación, ya que los aspectos de la tecnología descritos en el presente documento no están limitados a este respecto. En algunas realizaciones, el o los micrófonos 112 pueden incluir un micrófono para cada posición potencial de una persona que habla en el entorno 110. Por ejemplo, cuando el entorno 110 es un automóvil, el entorno 110 puede incluir un micrófono para cada uno de los asientos más en el automóvil.

En algunas realizaciones, el o los sensores 114 pueden incluir cualquier número y tipo de sensores de hardware adecuados configurados para detectar información sobre el entorno y/o personas que hablan en el entorno 110. Por ejemplo, el o los sensores 114 pueden incluir uno o más sensores (por ejemplo, uno o más sensores de presión, uno o más sensores de cámara para proporcionar datos ópticos, uno o más sensores de movimiento, uno o más sensores configurados para determinar si un cinturón de seguridad se ha abrochado, etc.) configurado para detectar una posición y/o identidad de una persona que habla. Como otro ejemplo, el o los sensores 114 pueden incluir uno o más sensores configurados para medir aspectos del entorno 110. Por ejemplo, cuando el entorno 110 es un vehículo, el o los sensores 114 pueden configurarse para medir la velocidad del vehículo, determinar si una o más ventanas y/o puertas del vehículo están abiertas, determinar si una persona que habla está usando una o más aplicaciones 116 reconocidas por voz c y/o cualquier otra información adecuada sobre el entorno 110.

En algunas realizaciones, la o las aplicaciones 116 incluyen una o más aplicaciones reconocidas por voz con las que una persona que habla en el entorno 110 puede interactuar hablando. Los ejemplos de aplicaciones reconocidas por voz incluyen, entre otras, una aplicación de navegación controlada por voz (por ejemplo, a través de la cual un usuario puede obtener indicaciones para llegar a un destino), una aplicación de telefonía controlada por voz (por ejemplo, a través de un usuario puede conducir llamadas telefónicas), cualquier aplicación configurada para realizar síntesis de texto a voz (TTS), una aplicación de entretenimiento controlada por voz (por ejemplo, a través de la cual un usuario puede ver uno o más programas de televisión, navegar por Internet, jugar videojuegos, comunicarse con uno o más usuarios, etc.), una aplicación de información del automóvil habilitada con voz, en realizaciones donde el entorno 110 es un automóvil, un sistema de comunicación en el automóvil (ICC) que permite a los usuarios en un vehículo comunicarse entre sí, en realizaciones donde El entorno 110 es un automóvil, y una aplicación habilitada con voz para controlar uno o más electrodomésticos, calor, aire acondicionado y/o iluminación, en realizaciones donde el entorno 110 es un hogar inteligente.

El sistema 100 incluye además un componente 120 inteligente de análisis de escena acústica que puede configurarse para obtener y analizar la entrada desde el entorno 110, incluida la entrada obtenida a través del o los micrófonos 112 y el o los sensores 114. Por ejemplo, el componente 120 inteligente de análisis de escena acústica puede configurarse para obtener una señal acústica generada por el o los micrófono 112 y realizar el procesamiento para determinar si la señal acústica incluye una palabra de activación para una cualquiera de la o las aplicaciones 116 reconocidas por voz.

Como se muestra en la realización de la figura 1, el componente 120 inteligente de análisis de escena acústica incluye el componente 122 de identificación de la persona que habla, el componente 124 de análisis acústico, el componente 126 de reconocimiento automático de voz (ASR)/comprensión del lenguaje natural (NLU) y la lógica 128 de control conjunto. Cada uno de los componentes 122, 124, 126 y la lógica 128 (y los componentes 132 y 134 tratados con más detalle a continuación) pueden implementarse en software (por ejemplo, utilizando instrucciones ejecutables por procesador), en hardware o como una combinación de software y hardware.

En algunas realizaciones, el componente 122 de identificación de la persona que habla puede configurarse para identificar una persona que habla de un enunciado basado, al menos en parte, en la información obtenida del entorno 110. En algunas realizaciones, el componente 122 de identificación de la persona que habla puede configurarse para obtener una o más señales acústicas del entorno 110, como resultado del enunciado de una persona que habla (por ejemplo, señales acústicas generadas por el o los micrófonos 112 en respuesta a la recepción de un enunciado de la persona que habla) y procesar las señales acústicas para identificar a la persona que habla. Este procesamiento puede realizarse de cualquier manera adecuada. Por ejemplo, el componente 122 de identificación de la persona que habla puede obtener una o más características del discurso (por ejemplo, una impresión de voz) a partir de la señal o señales acústicas obtenidas del entorno 110 y comparar las características del discurso obtenidas con las características del discurso almacenadas de las personas que hablan registradas con el sistema 100 (por ejemplo, personas que hablan inscritas) para determinar la identidad de la persona que habla. Las características del discurso de las personas que hablan registradas pueden almacenarse en cualquier medio o medio de almacenamiento no transitorio legible por ordenador adecuado y, por ejemplo, pueden almacenarse en el almacén 125 de datos mostrado en la figura 1. En algunas realizaciones, el componente 122 de identificación de la persona que habla puede configurarse para ayudar a registrar nuevas personas que hablan con el sistema 100 usando cualquier técnica de inscripción de personas que hablan adecuada. Por ejemplo, el componente 122 de identificación de la persona que habla puede configurarse para inscribir una persona que habla durante el tiempo de ejecución cuando las características del discurso de la persona que habla no coinciden con ninguna de las características del discurso (de otras personas que hablan) almacenadas por el sistema 100. En algunas realizaciones, el componente 122 de identificación de la persona que habla puede usar uno o más modelos estadísticos (por ejemplo, modelos estadísticos específicos de la persona que habla) que representan la biometría de voz de las personas que hablan registradas en el sistema 100 para determinar la identidad de la persona que habla. Sin embargo, debe apreciarse que el componente 122 de identificación de la persona que habla podría usar cualquier técnica de reconocimiento de la persona que habla adecuada para determinar la identidad de la persona que habla a partir de la o las señales acústicas obtenidas del entorno 110, ya que los aspectos de la tecnología descritos en el presente documento no están limitados en este sentido.

En algunas realizaciones, el componente 122 de identificación de la persona que habla puede configurarse para

identificar una persona que habla en función, al menos en parte, de información distinta de la información acústica obtenida del entorno de la persona que habla. Por ejemplo, el componente 122 puede obtener información sobre la posición de la persona que habla (por ejemplo, del componente 124 de análisis acústico, del o los sensores 114, etc.) y usar información sobre la posición de la persona que habla para determinar la identidad o probable identidad de la persona que habla. Por ejemplo, una persona puede ser, típicamente, el conductor de un vehículo y el componente 122 puede determinar la identidad o la probable identidad de esta persona que habla determinando que la entrada de voz se recibió desde el asiento del conductor del vehículo. Sin embargo, debe apreciarse que el componente 122 puede usar cualquier otra información adecuada para determinar la identidad o probable identidad de la persona que habla, ya que los aspectos de la tecnología descritos en el presente documento no están limitados en este sentido.

En algunas realizaciones, el componente 124 de análisis acústico puede configurarse para procesar cualquier señal acústica obtenida en el entorno 110 para obtener (por ejemplo, para detectar y/o estimar) diversas cantidades de interés sobre el entorno acústico de la persona que habla, ejemplos de cuyas cantidades se proporcionan más adelante. Sin embargo, estos ejemplos son ilustrativos y no limitativos, ya que el componente 124 puede configurarse para procesar señal o señales acústicas para obtener cualquier otra cantidad adecuada de interés sobre el entorno acústico del la persona que habla.

En algunas realizaciones, el componente 124 de análisis acústico puede configurarse para caracterizar cualquier ruido acústico presente en el entorno acústico de la persona que habla. Ejemplos no limitativos de dicho ruido acústico incluyen ruido ambiental (por ejemplo, debido al viento, lluvia, etc.), ruido eléctrico (por ejemplo, zumbido de un dispositivo eléctrico, zumbido de una línea eléctrica a 60Hz, etc.), música de fondo, e interferencia de voz por una o más personas que hablan (por ejemplo, ruido de balbuceo). El componente 124 puede configurarse para usar cualquier técnica de estimación de ruido adecuada para identificar la presencia de ruido acústico, determinar el tipo de ruido acústico presente y/o determinar la energía/potencia del ruido acústico presente (por ejemplo, en cualquier parte adecuada del espectro incluido en una o múltiples subbandas). Cualquier parte de esta u otra información determinada por el componente 124 sobre el ruido acústico puede ser utilizada por el sistema 100 en apoyo de varias tareas, incluida la eliminación del ruido de las señales acústicas obtenidas por el o los micrófonos 112 (por ejemplo, a través de una técnica e mejora de voz adecuada), estimar y eliminar componentes de eco que surgen en los micrófonos de la música reproducida (por ejemplo, mediante la cancelación de eco acústico), indicaciones de voz u otras señales conocidas internamente por el sistema, estimando la relación señal a ruido, realizando detección de actividad de voz, configuración parámetros de algoritmos de reconocimiento de voz (por ejemplo, para compensar y/o de otro modo dar cuenta de la presencia de ruido), determinando si los resultados del reconocimiento de voz deben procesarse posteriormente para tener en cuenta la presencia de ruido, y similares.

En algunas realizaciones, el componente 124 de análisis acústico puede configurarse para realizar detección de actividad de voz, a veces denominada detección de actividad de voz, para identificar partes de la señal acústica que probablemente contengan voz de una o múltiples personas que hablan. El componente 124 puede configurarse para realizar una detección de actividad de voz basada, al menos en parte, en una cantidad de energía/potencia detectada en la señal o señales acústicas por encima de la cantidad de ruido acústico determinado para estar presente en la señal o señales acústicas y/o de cualquier otra forma adecuada, ya que los aspectos de la tecnología descritos en el presente documento no están limitados en este sentido.

En algunas realizaciones, el componente 124 de análisis acústico puede configurarse para determinar la ubicación de una persona que habla en el entorno 110 en función de la señal o señales acústicas proporcionadas por el o los micrófonos 112. El componente 124 puede determinar la ubicación de la persona que habla en el entorno 110 aplicando cualquier técnica adecuada de localización de fuente acústica y/o formación de haz a la señal o señales acústicas. De forma adicional o alternativa, el componente 124 puede usar señales acústicas proporcionadas por múltiples micrófonos para reducir o eliminar el ruido acústico presente en las señales acústicas. Esto se puede hacer usando cualquier técnica adecuada de mejora de voz multimicrófono, que, por ejemplo, puede usar la formación de haces o aprovechar la correlación entre múltiples señales acústicas obtenidas por el o los micrófonos 112.

En algunas realizaciones, el componente 126 de ASR/NLU puede configurarse para realizar reconocimiento automático de voz y/o comprensión del lenguaje natural en las señales acústicas obtenidas en el entorno 110. El componente 126 de ASR/NLU puede incluir al menos un motor ASR configurado para realizar el reconocimiento de voz en la o las señales acústicas obtenidas en el entorno 110. El al menos un motor ASR puede configurarse para realizar reconocimiento automático de voz usando uno o más modelos acústicos, una o más gramáticas, uno o más transductores de estado finito, uno o más modelos de lenguaje, uno o más modelos de pronunciación, uno o más vocabularios, y/o cualquier otro componente adecuado para realizar ASR. El al menos un motor ASR puede configurarse para implementar cualquier técnica o técnicas ASR adecuadas, incluida cualquier técnica que haga uso de los componentes descritos anteriormente de un motor ASR, y puede incluir instrucciones ejecutables por el procesador que, cuando son ejecutadas por el sistema 100, realizan tales técnicas de ASR. El texto obtenido al reconocer la voz presente en la señal o señales acústicas obtenidas en el entorno 110 puede usarse para determinar si el discurso incluye alguna palabra de activación. Por ejemplo, el texto obtenido al reconocer el discurso de una persona que habla se puede comparar con las entradas en una lista de palabras de activación asociadas con la persona que habla. El componente 126 de ASR/NLU puede incluir al menos un motor de NLU configurado para

realizar una o más técnicas de NLU para deducir la intención de una persona que habla (por ejemplo, para determinar una acción que la persona que habla desea realizar) y puede incluir instrucciones ejecutables por el procesador que, cuando son ejecutadas por el sistema 100, realizan tales técnicas de NLU.

En algunas realizaciones, el almacén 125 de datos puede configurarse para almacenar información sobre una o más personas que hablan registradas (por ejemplo, inscritas) con el sistema 100. Por ejemplo, el almacén 125 de datos puede almacenar información sobre la identidad de una persona que habla, tal como el nombre de la persona que habla y/u otra información que especifique la identidad de la persona que habla. Como otro ejemplo, el almacén 125 de datos puede almacenar una o más listas de una o más palabras de activación asociadas con la persona que habla. La lista o listas pueden indicar, para una palabra de activación en particular, la o las aplicaciones controladas por voz para las que la palabra de activación puede usarse como disparador de voz. Como otro ejemplo más, el almacén 125 de datos puede almacenar información de control de acceso asociada con una persona que habla (por ejemplo, información que indica qué aplicaciones controladas por voz le está permitido controlar o no a la persona que habla. Como otro ejemplo más, el almacén 125 de datos puede almacenar información sobre el comportamiento de una persona que habla, que incluye, entre otros, información que indica una o más aplicaciones controladas por voz a las que la persona que habla accedió anteriormente, información que indica dónde se encuentra normalmente una persona que habla cuando emite comandos (por ejemplo, asiento del conductor) e información que indica las preferencias y/o intereses de la persona que habla (por ejemplo, el programa de radio favorito de la persona que habla, el canal de televisión, el género musical, etc.). Como otro ejemplo más, el almacén 125 de datos puede almacenar información sobre una persona que habla que se puede usar para adaptar un reconocedor de voz a la persona que habla (por ejemplo, uno o más enunciados de inscripción). De forma adicional o alternativa, el almacén 125 de datos puede configurarse para almacenar cualquier otra información adecuada que pueda ser utilizada por el sistema 100 para realizar la detección de palabras de activación. Por ejemplo, el almacén 125 de datos puede configurarse para almacenar información obtenida de uno o más sensores (por ejemplo, los sensores 114 descritos anteriormente).

En algunas realizaciones, el almacén 125 de datos puede organizar al menos algunos de los datos en múltiples registros de datos. Puede haber cualquier número adecuado de registros de datos en el almacén 125 de datos y pueden formatearse de cualquier manera adecuada. Un registro de datos puede incluir información asociada con una persona que habla que incluye, por ejemplo, al menos algunos (por ejemplo, todos) de los tipos de información descritos anteriormente. Por ejemplo, como se muestra en la figura 4, el almacén 125 de datos puede almacenar múltiples registros de datos, cada uno de los cuales incluye información que identifica a una persona que habla, una o múltiples posiciones en un vehículo que se sabe que la persona que habla ha ocupado previamente, una lista de aplicaciones controladas por voz que la persona que habla está autorizada para disparar usando una palabra de activación y una lista de una o más palabras de activación asociadas con la persona que habla. Por ejemplo, se sabe que la persona que habla "Alice" ocupó previamente el asiento del conductor y el asiento del lado del pasajero delantero. Alice está autorizada para disparar todas las aplicaciones controladas por voz utilizando la palabra de activación "¡Oye!". Como otro ejemplo, se sabe que la persona que habla "Charlie" ocupó previamente el asiento trasero y el asiento del lado del pasajero delantero. Charlie está autorizado para disparar solo las aplicaciones de entretenimiento y telefonía, pero no cualquier otra aplicación (por ejemplo, al no ser Charlie el conductor del coche, no puede activar la aplicación de navegación controlada por voz, mientras que Alice y Bob pueden activar tal aplicación porque han conducido el automóvil). Charlie puede activar una aplicación controlada por voz para entretenimiento usando la palabra de activación "TV" y la aplicación de telefonía mediante la palabra de activación "Llamar a alguien". David no está autorizado para activar verbalmente ninguna aplicación controlada por voz y el almacén 125 de datos no almacena ninguna palabra de activación personalizada para David. Debe apreciarse que, aunque, en algunas realizaciones, el almacén 125 de datos puede organizar al menos algunos de los datos utilizando registros de datos, los aspectos de la tecnología descritos en el presente documento no están limitados en este sentido. El almacén 125 de datos puede almacenar datos en una o más bases de datos de cualquier tipo adecuado, uno o más archivos, una o más tablas, utilizando cualquier estructura o esquemas de indexación adecuados.

En algunas realizaciones, la lógica 128 de control conjunto puede configurarse para recopilar información obtenida del entorno 110 y/o uno o más de otros componentes del sistema 100 y procesar la información recopilada de modo que pueda usarse actualmente o en el futuro para fomentar aún más una obtenerse más tareas realizadas por el sistema 100 como, por ejemplo, detección de palabras de activación y/o control de acceso. En algunas realizaciones, la lógica 128 de control conjunto puede configurarse para organizar la información obtenida y almacenar la información organizada (por ejemplo, en el almacén 125 de datos) y/o proporcionar la información organizada a uno o más componentes del sistema 100 (por ejemplo, activar el componente 132 de detección de palabras y el componente 134 de control de acceso).

En algunas realizaciones, la lógica 128 de control conjunto puede obtener, de una o múltiples fuentes, diversos tipos de información relacionada con un enunciado pronunciado por una persona que habla en el entorno 110 y almacenar la información organizada (por ejemplo, en el almacén 125 de datos) y/o proporcionar la información organizada a uno o más componentes del sistema 100 (por ejemplo, el componente 132 de detección de palabras de activación y/o el componente 134 de control de acceso). Por ejemplo, la lógica 128 de control conjunto puede obtener, para un enunciado pronunciado en particular, información que incluye la identidad o probable identidad de la persona que

habla del enunciado pronunciado (por ejemplo, del componente 122 de identificación de la persona que habla), una posición de la persona que habla en el entorno 110 (por ejemplo, del componente 124 de análisis acústico), el texto correspondiente a un resultado de realizar ASR en el enunciado pronunciado (por ejemplo, del módulo 126 de ASR/NLU), información que indica cuáles de las aplicaciones 116 controladas por voz se están ejecutando, información asociada con la persona que pronuncia el enunciado (por ejemplo, del almacén 125 de datos), información que indica qué aplicación controlada por voz está intentando activar la persona que habla, y/o cualquier otra información adecuada relacionada con el enunciado. La información asociada con la persona que habla del enunciado puede incluir información que indica una o más palabras de activación asociadas con la persona que habla, información que indica qué aplicaciones reconocidas por voz está permitido que la persona que habla controle o no, las preferencias y/o intereses de la persona que habla, y/o cualquier otra información adecuada asociada con la persona que habla.

En consecuencia, en algunas realizaciones, la información obtenida por la lógica 128 de control conjunto puede usarse para actualizar el contenido del almacén 125 de datos. En algunas realizaciones, la lógica 128 de control conjunto puede actualizar el contenido del almacén 125 de datos en tiempo real con la información que obtiene del entorno 110 y/o uno o más componentes del sistema 100.

En algunas realizaciones, el componente 132 de palabra de activación puede configurarse para determinar, basándose al menos en parte en la información proporcionada por la lógica 128 de control conjunto, si una persona que habla pronunció una palabra de activación para cualquiera de la o las aplicaciones 116 controladas por voz. Por ejemplo, el componente 132 de palabra de activación puede determinar si la persona que habla pronunció una palabra de activación comparando los resultados de realizar el reconocimiento de voz automático en el enunciado de la persona que habla con las palabras de activación en una lista de palabras de activación asociadas con la persona que habla. De forma adicional o alternativa, el componente de palabra de activación puede determinar si una persona que habla en una posición particular pronunció una palabra de activación comparando los resultados de realizar ASR en el enunciado de la persona que habla con palabras de activación asociadas con cualquier persona de activación que pueda estar en la posición en particular. Por ejemplo, incluso si en algún caso el sistema 100 no ha determinado la identidad de la persona que habla, el sistema puede haber determinado la posición de la persona que habla (por ejemplo, el asiento del conductor de un automóvil) y puede tener información que indique qué persona que habla registrada en el sistema se ha sentado previamente en el asiento del conductor. El componente 132 de detección de palabras de activación puede comparar los resultados de reconocer el enunciado con las palabras de activación asociadas con cualquier persona que habla registrada con el sistema 100 que se ha sentado previamente en el asiento del conductor. Sin embargo, debe apreciarse que el componente 132 de la palabra de activación puede configurarse para determinar si una persona que habla pronunció una palabra de activación de cualquier otra manera adecuada en función de la información disponible en el sistema 100 (por ejemplo, información obtenida por la lógica 128 de control conjunto), ya que los aspectos de la tecnología descritos en el presente documento no están limitados en este sentido.

En algunas realizaciones, el componente 134 de control de acceso puede configurarse para determinar si una persona que habla que ha pronunciado una palabra de activación para una aplicación controlada por voz está autorizada para activar verbalmente la aplicación controlada por voz. El componente 134 de control de acceso puede hacer esta determinación basándose en la información obtenida de la lógica 128 de control conjunto. Por ejemplo, el componente 134 de control de acceso puede obtener, a partir de la lógica 128 de control conjunto, información que indica la identidad de una persona que habla, la posición de la persona que habla y/o la aplicación controlada por voz que la persona que habla está intentando activar. El módulo 134 de control de acceso también puede obtener información que indica qué aplicaciones reconocidas por voz puede activar la persona que habla identificada y/o qué aplicaciones reconocidas por voz pueden activarse desde la posición de la persona que habla. Según la información, el componente 134 de control de acceso puede determinar si la persona que habla puede activar verbalmente la aplicación controlada por voz desde la posición en la que estaba hablando. Cuando el componente 134 determina que la persona que habla puede controlar la aplicación controlada por voz, el componente 134 puede otorgarle a la persona que habla acceso a la aplicación controlada por voz. Por otro lado, cuando el componente 134 determina que la persona que habla no puede controlar la aplicación controlada por voz, el componente 134 puede limitar (por ejemplo, no permitir) el acceso de la persona que habla a la aplicación controlada por voz. Como un ejemplo, cuando un pasajero del asiento trasero desea verificar la información sobre un automóvil (por ejemplo, para determinar la presión de los neumáticos del automóvil, la cantidad de gasolina en el automóvil, el kilometraje del automóvil, la velocidad del automóvil, la temperatura en el automóvil, etc.) y pronuncia la palabra de activación "Hola, sistema del coche" para activar verbalmente la aplicación controlada por voz para proporcionar información sobre el automóvil, el componente 134 de acceso puede obtener información que indica que los pasajeros en el asiento trasero del automóvil no pueden interactuar con esta aplicación controlada por voz (y/o información que indica que la persona que habla en particular no puede activar verbalmente esta aplicación controlada por voz) y no permite que la persona que habla controle la aplicación controlada por voz para proporcionar información sobre el automóvil.

Debe apreciarse que el sistema 100 controlado por voz puede configurarse para procesar la voz provista simultáneamente por múltiples personas que hablan que pueden tratar de participar y/o interactuar con múltiples aplicaciones controladas por voz en el mismo entorno acústico. Como ejemplo ilustrativo, un conductor, un pasajero delantero y un pasajero trasero (por ejemplo, detrás del conductor) pueden tener una conversación en un automóvil.

El conductor desea controlar la aplicación de navegación controlada por voz para cambiar el destino de navegación y pronuncia una palabra de activación para activar verbalmente la aplicación de navegación. Al mismo tiempo que el conductor pronuncia una palabra de activación, el pasajero delantero puede estar hablando con los otros pasajeros. La señal de voz del pasajero delantero puede transmitirse a través del sistema de comunicación en el automóvil (ICC) a los otros pasajeros. En este escenario, el sistema 100 puede procesar las señales de voz detectadas en el automóvil para determinar que el conductor ha emitido una palabra de activación para la aplicación de navegación (por ejemplo, determinando la identidad del conductor, reconociendo la voz del conductor y comparando la voz reconocida contra palabras de activación en una lista de palabras de activación asociadas con el conductor) y permita que el conductor controle verbalmente la aplicación de navegación dirigiendo el habla detectada por el micrófono del conductor a la aplicación de navegación controlada por voz y excluyendo que la voz detectada por el micrófono del conductor se proporcione al sistema ICC. Además, el componente 120 puede permitir que el pasajero delantero y el pasajero trasero conversen utilizando el sistema ICC, pero puede evitar que se proporcione la voz detectada por sus micrófonos a la aplicación de navegación controlada por voz con la cual el conductor está interactuando.

Como otro ejemplo, el conductor de un automóvil desea iniciar una llamada de teleconferencia con el pasajero delantero del automóvil y una persona remota y pronuncia una palabra de activación "Hola sistema telefónico" para activar verbalmente una aplicación de telefonía. Aproximadamente al mismo tiempo, el pasajero del asiento trasero desea cambiar el canal de televisión y pronuncia la palabra de activación "Hola, sistema de televisión". En este escenario, el sistema 100 puede procesar las señales de voz detectadas en el automóvil para determinar que el conductor ha emitido una palabra de activación para la aplicación de telefonía (por ejemplo, determinando la identidad del conductor, la posición del conductor y consultando una lista personalizada de palabras de activación asociadas con la persona que habla) y puede dirigir el habla detectada por el micrófono del conductor a la aplicación de telefonía. De manera similar, el sistema 100 puede procesar las señales de voz detectadas en el automóvil para determinar que el pasajero del asiento trasero ha emitido una palabra de activación para el programa de entretenimiento habilitado para el habla, y puede dirigir la voz detectada por el micrófono del pasajero del asiento trasero.

Debe apreciarse que el sistema 100 es ilustrativo y que hay variaciones del sistema 100. Por ejemplo, en algunas realizaciones, los componentes 132 y 134 pueden ser parte del mismo componente y/o pueden ser parte del componente 120 de análisis de escena acústica inteligente. Más en general, las funciones realizadas por los componentes ilustrados en la realización de la figura 1 pueden realizarse mediante uno o más de otros componentes en otras realizaciones. También debe apreciarse que el sistema 100 puede tener uno o más componentes adicionales además del componente ilustrado en la figura 1.

La figura 2 es un diagrama de bloques de otro sistema 200 ilustrativo controlado por voz, de acuerdo con algunas realizaciones de la tecnología descrita en el presente documento. El sistema 200 ilustrativo es parte del vehículo 202 e incluye los micrófonos 206a, 206b, 206c y 206d configurados para detectar el enunciado pronunciado por las personas que hablan 204a, 204b, 204c y 204d. Cada uno de los micrófonos 206a-d puede configurarse para detectar el enunciado de cualquiera de las personas que hablan 204a-d. Cada uno de los micrófonos 206a-d puede ser de cualquier tipo adecuado, cuyos ejemplos se proporcionan en el presente documento. El sistema 200 también incluye el sensor 205 configurado para detectar si el conductor del automóvil está sentado. El sensor 205 puede ser un sensor de presión, un sensor del cinturón de seguridad y/o cualquier otro sensor adecuado configurado para detectar el presente de un conductor. En otras realizaciones, el sistema 200 puede incluir uno o más de otros sensores configurados para detectar la presencia de uno o más pasajeros en el automóvil 202, pero estos sensores no se muestran en el presente documento por claridad de presentación (y no a modo de limitación).

En la realización ilustrada, el sistema 200 incluye al menos un procesador de hardware de ordenador (por ejemplo, al menos un ordenador) 210 configurado para ejecutar múltiples aplicaciones reconocidas por voz, que incluyen, entre otras, la aplicación 212a de telefonía manos libres, la aplicación 212b de comunicación en el automóvil, y una aplicación 212c controlada por voz configurada para soportar un diálogo con una persona que habla al menos en parte mediante el uso de técnicas de síntesis de voz. El sistema 200 se puede configurar para recibir enunciados de voz pronunciados por personas que hablan 204a-d y se puede configurar para determinar si alguno de los enunciados de voz detectados incluye una palabra de activación para una de las aplicaciones 212a-c y/o si la persona que habla de una palabra de activación en particular para una aplicación particular está autorizado para controlar la aplicación controlada por voz en particular.

Para procesar la o las señales acústicas detectadas en el automóvil 202 por el o los micrófono 206a-d y cualquier otro sensor (por ejemplo, el sensor 205), el al menos un procesador 210 puede estar configurado para ejecutar un componente de análisis de escena acústica 214, un componente de detección de palabras de activación 216 y el componente 218 de control de acceso. El componente 214 de análisis de escena acústica puede configurarse para obtener y analizar la entrada del automóvil 202, incluida la entrada obtenida por los micrófonos 206a-d y el sensor 205. Por ejemplo, el componente 214 puede configurarse para obtener una señal acústica generada por uno de los micrófonos 206a-d y realizar el procesamiento para determinar si la señal acústica incluye una palabra de activación para cualquiera de las aplicaciones 212 (a) - (c) controladas por voz. El componente 214 del análisis de escenas acústicas puede configurarse para operar de cualquiera de las formas descritas con referencia al componente 120

del análisis de escenas acústicas descrito con referencia a la figura 1 y, en algunas realizaciones, puede incluir uno o más componentes (por ejemplo, un componente de identificación de la persona que habla, un componente de análisis acústico, un componente ASR/NLU, lógica de control conjunto, etc.) descritos con referencia a la figura 1.

El componente 216 de detección de palabras de activación puede configurarse para determinar, en función de, al menos en parte, la información proporcionada por el componente 214 de análisis de escena acústica, si una persona que habla pronunció una palabra de activación para cualquiera de las aplicaciones 212a-c controladas por voz y puede funcionar de cualquiera de las formas descritas con referencia al componente 132 de detección de palabras de activación descrito con referencia a la figura 1.

El componente 218 de control de acceso puede configurarse para determinar si una persona que habla que ha pronunciado una palabra de activación para una aplicación controlada por voz está autorizada para activar verbalmente la aplicación controlada por voz. El componente 218 de control de acceso puede hacer esta determinación basándose al menos en parte en la información obtenida del componente 214 de análisis de escena acústica y puede operar en cualquiera de las formas descritas con referencia al componente 134 de control de acceso descrito con referencia a la figura 1.

Debe apreciarse que el sistema 200 es ilustrativo y que hay variaciones del sistema 200. Por ejemplo, las funciones realizadas por los componentes ilustrados en la realización de la figura 2 pueden realizarse mediante uno o más de otros componentes en otras realizaciones. También debe apreciarse que el sistema 200 puede tener uno o más componentes adicionales además del componente ilustrado en la figura 2.

La figura 3 es un diagrama de flujo de un proceso 300 ilustrativo para detectar una palabra de activación en un enunciado basado, al menos en parte, en información indicativa de la identidad de la persona que habla del enunciado, de acuerdo con algunas realizaciones de la tecnología descrita en el presente documento. El proceso 300 puede realizarse mediante cualquier sistema adecuado para detectar una palabra de activación en un enunciado y, por ejemplo, puede realizarse mediante el sistema 100 reconocido por voz descrito con referencia a la figura 1 o mediante el sistema 200 reconocido por voz descrito con referencia a la figura 2.

El proceso 300 comienza en el acto 302, donde se recibe una señal acústica que contiene un enunciado hablado por una persona que habla. La señal acústica puede ser generada por un micrófono al menos en parte como resultado de recibir y/o detectar el enunciado hablado por la persona que habla. Por ejemplo, después de que una persona que habla pronuncia una palabra de activación para una aplicación controlada por voz en particular, un micrófono puede detectar el enunciado y generar una señal acústica basada en el enunciado detectado, cuya señal acústica puede recibirse en el acto 302. En algunas realizaciones, múltiples micrófonos pueden recibir y/o detectar un enunciado hablado por una persona que habla y las señales acústicas generadas por los múltiples micrófonos pueden recibirse en el acto 302.

El siguiente proceso 300 procede al acto 304, donde se obtiene información indicativa de la identidad de la persona que habla. En algunas realizaciones, obtener información indicativa de la identidad de la persona que habla puede incluir recibir información que especifica la identidad de la persona que habla. En algunas realizaciones, obtener información indicativa de la identidad de la persona que habla puede incluir procesar información indicativa de la identidad de la persona que habla para determinar la identidad de la persona que habla.

En algunas realizaciones, por ejemplo, obtener información indicativa de la identidad de la persona que habla comprende procesar la o las señales acústicas obtenidas en el acto 302 (por ejemplo, usando biometría de voz) con el fin de determinar la identidad de la persona que habla. Por ejemplo, el proceso 300 de ejecución del sistema puede obtener una o más características del discurso (por ejemplo, una impresión de voz) de la o las señales acústicas recibidas en el acto 302 y comparar las características del discurso obtenidas contra las características del discurso almacenadas para cada una de las múltiples personas que hablan registradas con el sistema para determinar la identidad de una persona que habla. Sin embargo, debe apreciarse que se puede usar cualquier técnica(s) de reconocimiento de la persona que habla adecuada como parte del acto 304 para determinar la identidad de una persona que habla a partir de las señales acústicas) recibidas en el acto 302, ya que los aspectos de la tecnología descrita en el presente documento no están limitados a este respecto.

En algunas realizaciones, la obtención de información indicativa de la identidad de la persona que habla comprende determinar la posición de la persona que habla en el entorno acústico. Tal determinación puede hacerse a partir de los datos recopilados por uno o más micrófonos y/o uno o más de otros sensores en el entorno acústico del la persona que habla. Por ejemplo, cuando varios micrófonos detectan el habla de una persona que habla, las señales detectadas por los micrófonos se pueden usar para determinar la ubicación de la persona que habla (por ejemplo, utilizando técnicas de formación de haz). Como otro ejemplo, cuando una persona que habla se encuentra en un vehículo (por ejemplo, un automóvil), la posición del la persona que habla se puede determinar al menos en parte mediante el uso de uno o más sensores de presión (por ejemplo, en un asiento, en el cinturón de seguridad, etc.) y/u otros sensores (por ejemplo, una cámara de vídeo). Por lo tanto, la posición de una persona que habla se puede determinar utilizando uno o más sensores de cualquier tipo adecuado, incluidos, pero sin limitación, uno o más sensores acústicos, uno o más sensores de presión y una o más cámaras (por ejemplo, una o más cámaras de

vídeo).

En algunas realizaciones, la posición de una persona que habla en el entorno acústico puede usarse para deducir la identidad probable de la persona que habla o para deducir identidades de múltiples personas que hablan, una de las cuales es probable que esté hablando. En algunos casos, la posición de la persona que habla puede usarse para identificar una sola persona que habla probable. Por ejemplo, un proceso 300 de ejecución del sistema controlado por voz puede determinar que una persona que habla está sentada en el asiento delantero de un vehículo y, según la información que indica que la persona que habla "S" es el conductor más frecuente del vehículo, determinar que "S" es la probable persona que habla. En otros casos, la posición de la persona que habla se puede usar para identificar varias personas que hablan de las cuales es probable que esté hablando (por ejemplo, cuando se determina que la persona que habla en un automóvil está sentada en el asiento del conductor y el proceso 300 de ejecución del sistema controlado por voz conoce los múltiples adultos que pueden conducir un automóvil) entre todas las posibles personas que hablan. En algunos casos, la posición determinada por la persona que habla se puede usar para determinar qué personas que hablan es menos probable que estén hablando (por ejemplo, el sistema reconocido por voz puede determinar que aunque los niños registrados con el sistema reconocido por voz, es probable que los niños no estén hablando porque no conducen y se determina que la persona que habla está sentada en el asiento del conductor).

A continuación, el proceso 300 pasa a actuar 306, donde se determina si el enunciado recibido en el acto 302 incluye una palabra de activación designada para cualquier aplicación(es) controlada por voz. Esta determinación puede hacerse basándose, al menos en parte, en la información indicativa de la identidad de la persona que habla obtenida en el acto 304. En algunas realizaciones, la información indicativa de la identidad de la persona que habla puede especificar la identidad de la persona que habla y/o procesarse para determinar la identidad de la persona que habla, y la identidad de la persona que habla puede usarse para determinar si el enunciado recibido en el acto 302 incluye una palabra de activación designada para cualquier aplicación controlada por voz. La identidad de la persona que habla puede usarse para hacer esta determinación de cualquier manera adecuada, cuyos ejemplos se describen en el presente documento.

En algunas realizaciones, por ejemplo, la identidad de la persona que habla puede usarse para acceder a una lista personalizada de una o más palabras de activación asociadas con la persona que habla identificada. El contenido de la lista a la que se accede, a su vez, se puede usar para determinar si el enunciado recibido en el acto 302 incluye una palabra de activación designada para cualquier aplicación controlada por voz. En algunas realizaciones, la lista de una o más palabras de activación asociadas con la persona que habla identificada se puede proporcionar a un reconocedor de voz (que, por ejemplo, puede ser parte del componente 126 de ASR/NLU que se muestra en la figura 1) para su uso en realizar un reconocimiento de voz automático en el enunciado de la persona que habla para determinar si el enunciado contiene alguna palabra de activación en la lista de palabras de activación accedidas. El reconocedor de voz puede usar la lista de palabras de activación para facilitar el reconocimiento de la palabra de activación designada de cualquier manera adecuada, incluyendo, pero sin limitación, restringir su espacio de búsqueda a las palabras en la lista de palabras de activación, utilizando una gramática basada en la lista de palabras de activación y/o adaptando uno o más de sus componentes (por ejemplo, uno o más modelos acústicos, uno o más modelos de pronunciación, uno o más modelos de lenguaje, uno o más transductores de estado finito, uno o más léxicos o vocabularios, y/o cualquier otro componente adecuado en función dela lista de palabras de activación.

De forma adicional o alternativa, la lista de palabras de activación asociadas con la persona que habla se puede utilizar para compensar la interferencia presente en la o las señales acústicas recibidas en el acto 302, lo que a su vez facilita determinar con precisión si el enunciado incluye o no la palabra de activación designada para cualquier aplicación controlada por voz. Por ejemplo, cuando un reconocedor de voz reconoce solo una parte de una palabra de activación en particular, la parte reconocida se puede comparar con una o más entradas en la lista de palabras de activación para identificar que la palabra de activación designada fue probablemente pronunciada por la persona que habla a pesar de que no se había reconocido la palabra de activación completa debido a la presencia de interferencias. Como ejemplo específico, una persona que habla puede pronunciar la frase "Buenos días coche" como palabra de activación para una aplicación controlada por voz en un automóvil. Sin embargo, debido a la interferencia causada por una o más personas que hablan al mismo tiempo, el proceso 300 de ejecución del sistema controlado por voz puede haber reconocido solo la parte de "días coche" del enunciado. Al comparar la parte detectada "días coche" con la entrada "Buenos días coche" en la lista de palabras de activación asociadas con la persona que habla, el sistema controlado por voz puede determinar que, porque la parte reconocida coincide al menos parcialmente con una palabra de activación en la lista, es probable que la persona que habla pronunciara la palabra de activación "Buenos días coche" y puede permitir que la persona que habla controle verbalmente la aplicación controlada por voz.

Cuando se determina en el acto 306 que la emisión de la persona que habla no incluye una palabra de activación designada para ninguna aplicación controlada por voz, el proceso 300 se completa. Por otro lado, cuando se determina en el acto 308 que el enunciado de la persona que habla incluye una palabra de activación designada para una aplicación controlada por voz en particular, el proceso 300 pasa al bloque 308 de decisión, donde se determina si la persona que habla está autorizada para controlar la aplicación controlada por voz en particular para la cual la persona que habla pronunció una palabra de activación. Esta determinación puede hacerse de cualquier

manera adecuada. Por ejemplo, el proceso 300 de ejecución del sistema puede acceder a información que indica qué personas que hablan están (y/o no) autorizadas para controlar la aplicación controlada por voz en particular. Cuando la información a la que se accede indica que la persona que habla está autorizada para controlar la aplicación en particular, se puede determinar en el bloque 308 de decisión que la persona que habla está autorizada para controlar la aplicación controlada por voz en particular. Por otro lado, cuando la información a la que se accede indica que la persona que habla no está autorizada para controlar la aplicación en particular, se puede determinar en el bloque 308 de decisión que la persona que habla no está autorizada para controlar la aplicación controlada por voz en particular. Como otro ejemplo, el proceso 300 de ejecución del sistema puede acceder a ka información que indica que las personas que hablan en ciertas posiciones en el entorno acústico están (y/o no) autorizadas para controlar la aplicación particular controlada por voz (por ejemplo, las personas que hablan en el asiento trasero de un coche no están autorizadas para controlar la aplicación de navegación). En este caso, la posición de la persona que habla (que puede obtenerse en el acto 304 del proceso 300) puede usarse para determinar si la persona que habla está autorizada para controlar la aplicación controlada por voz.

Cuando se determina en el bloque 308 de decisión que la persona que habla está autorizada para controlar la aplicación controlada por voz en particular, el proceso 300 pasa al acto 310 donde la persona que habla puede controlar la aplicación controlada por voz, por ejemplo, proporcionando uno o más comandos de voz. Por otro lado, cuando se determina en el bloque 308 de decisión que la persona que habla no está autorizada para controlar la aplicación controlada por voz en particular, el proceso 300 pasa al acto 312 donde el proceso 300 de ejecución del sistema controlado por voz requiere una o más etapas para no permitir que la persona que habla controle la aplicación controlada por voz en particular.

Debe apreciarse que el proceso 300 es ilustrativo y que son posibles variaciones del proceso 300. Por ejemplo, aunque en la realización ilustrada, el proceso 300 incluye los actos 308, 310 y 312 relacionados con la funcionalidad de control de acceso, en otras realizaciones, los actos 308, 310 y 312 pueden omitirse o pueden ser opcionales.

Una implementación ilustrativa de un sistema informático 500 que se puede usar en conexión con cualquiera de las realizaciones de la divulgación proporcionada en el presente documento se muestra en la figura 5. El sistema informático 500 puede incluir uno o más procesadores 510 y uno o más artículos de fabricación que comprenden medios de almacenamiento no transitorios legibles por ordenador (por ejemplo, memoria 520 y uno o más medios 530 de almacenamiento no volátil). El procesador 510 puede controlar los datos escritos hacia y los datos de lectura desde la memoria 520 y el dispositivo 530 de almacenamiento no volátil de cualquier forma, dado que los aspectos de la divulgación proporcionados en el presente documento no están limitados a este respecto. Para realizar cualquiera de las funcionalidades descritas en el presente documento, el procesador 510 puede ejecutar una o más instrucciones ejecutables por el procesador almacenadas en uno o más medios de almacenamiento no transitorios legibles por ordenador (por ejemplo, la memoria 520), que pueden servir como medios de almacenamiento no transitorios legibles por ordenador que almacenan las instrucciones ejecutables por procesador para la ejecución mediante el procesador 510.

Los términos "programa" o "software" se usan en el presente documento en sentido general para hacer referencia a cualquier tipo de código informático o conjunto de instrucciones ejecutables por procesador que se pueden usar para programar un ordenador u otro procesador para implementar varios aspectos de las realizaciones, como se ha tratado anteriormente. Adicionalmente, debe apreciarse que, de acuerdo con un aspecto, uno o más programas de ordenador que, cuando se ejecutan, realizan métodos de la divulgación proporcionada en el presente documento, no tienen que residir en un solo ordenador o procesador, pero se pueden distribuir de forma modular entre diferentes ordenadores o procesadores para implementar varios aspectos de la divulgación proporcionada en el presente documento.

Las instrucciones ejecutables por el procesador pueden estar en muchas formas, tales como módulos de programa, ejecutadas por uno o más ordenadores u otros dispositivos. En general, los módulos de programa incluyen rutinas, programas, objetos, componentes, estructuras de datos, etc., que realizan tareas particulares o implementan tipos de datos abstractos particulares. De forma típica, la funcionalidad de los módulos de programa puede combinarse o distribuirse según se desee en diversas realizaciones.

Asimismo, las estructuras de datos se pueden almacenar en uno o más medios de almacenamiento no transitorios legibles por ordenador en cualquier forma adecuada. En aras de la simplicidad de la ilustración, las estructuras de datos se pueden mostrar de modo que tengan campos que están relacionados mediante la localización en la estructura de datos. Dichas relaciones pueden, asimismo, lograrse mediante la asignación de almacenamiento para los campos con localizaciones en un medio no transitorio legible por ordenador que transportan la relación entre los campos. Sin embargo, se puede usar cualquier mecanismo adecuado para establecer relaciones entre la información en campos de una estructura de datos, incluido mediante el uso de apuntadores, marcadores u otros mecanismos que establecen relaciones entre los elementos de datos.

Además, varios conceptos de la invención pueden incorporarse como uno o más procesos, de los cuales se han proporcionado ejemplos. Los actos realizados como parte de cada proceso pueden ordenarse de cualquier manera adecuada. En consecuencia, se pueden construir realizaciones en las que los actos se realicen en un orden

diferente al ilustrado, lo que puede incluir realizar algunos actos simultáneamente, aunque se muestren como actos secuenciales en realizaciones ilustrativas.

Debe entenderse que todas las definiciones, como se definen y usan en el presente documento, controlan a través de definiciones del diccionario y/o significados ordinarios de los términos definidos.

Como se usa en el presente documento, en la especificación y en las reivindicaciones, la frase "al menos uno", en referencia a una lista de uno o más elementos, debe entenderse que significa al menos un elemento seleccionado de uno o más de los elementos en la lista de elementos, pero no necesariamente incluye al menos uno de todos y cada uno de los elementos enumerados específicamente en la lista de elementos y no excluye ninguna combinación de elementos en la lista de elementos. Esta definición también permite que los elementos puedan estar presentes opcionalmente además de los elementos específicamente identificados dentro de la lista de elementos a los que se refiere la frase "al menos uno", ya sea relacionado o no con esos elementos específicamente identificados. Por lo tanto, como ejemplo no limitativo, "al menos uno de A y B" (o, e forma equivalente, "al menos uno de A o B" o, de forma equivalente "al menos uno de A y/o B") puede hacer referencia, en una realización, a al menos uno, que incluye opcionalmente más de uno, A, sin B presente (y opcionalmente incluyendo elementos distintos de B); en otra realización, al menos a uno, que incluye opcionalmente más de uno, B, sin A presente (y opcionalmente que incluye elementos distintos de A en otra realización más, a al menos uno, que incluye opcionalmente más de uno, A y al menos uno, que incluye opcionalmente más de uno, B (y opcionalmente que incluye otros elementos); etc.

La frase "y/o", tal como se usa en el presente documento en la memoria descriptiva y en las reivindicaciones, debe entenderse que significa "cualquiera o ambos" de los elementos así unidos, es decir, elementos que están presentes de manera conjunta en algunos casos y presentes de manera disjunta en otros casos. Los elementos múltiples enumerados con "y/o" deben construirse de la misma manera, es decir, "uno o más" de los elementos así unidos. Opcionalmente puede haber presentes otros elementos distintos de los elementos específicamente identificados por la cláusula "y/o", ya estén relacionados o no relacionados con los elementos específicamente identificados. Por lo tanto, como ejemplo no limitativo, una referencia a "A y/o B", cuando se usa junto con un lenguaje abierto como "que comprende", puede referirse, en una realización, a A solamente (opcionalmente incluyendo elementos distintos de B); en otra realización, a B solamente (incluyendo opcionalmente elementos distintos de A); en aún otra realización, a A y B (que incluyen opcionalmente otros elementos); etc.

El uso de términos ordinales como "primero", "segundo", "tercero", etc., en las reivindicaciones para modificar un elemento de la reivindicación no connota por sí mismo ninguna prioridad, precedencia u orden de un elemento de la reivindicación sobre otro o el orden temporal en que se realizan los actos de un método. Estos diez se usan simplemente como etiquetas para distinguir un elemento de la reivindicación que tiene un nombre determinado de otro elemento que tiene el mismo nombre (pero para el uso del término ordinal).

La fraseología y la terminología utilizadas en el presente documento son para fines de descripción y no deben considerarse limitativas. El uso de "que incluye", "que comprende", "que tiene", "que contiene", "que implica" y variaciones de los mismos quiere decir que abarca los elementos enumerados a continuación y elementos adicionales.

Habiendo descrito varias realizaciones de las técnicas descritas con detalle en el presente documento, los expertos en la técnica idearán fácilmente varias modificaciones y mejoras. En consecuencia, la descripción anterior es solo a modo de ejemplo, y no pretende ser limitativa. Las técnicas están limitadas solo como se define en las siguientes reivindicaciones.

REIVINDICACIONES

1. Un sistema para detectar al menos una palabra de activación designada para al menos una aplicación controlada por voz, comprendiendo el sistema:

al menos un micrófono; y
al menos un procesador de hardware de ordenador configurado para realizar:

recibir una señal acústica generada por el al menos un micrófono, al menos en parte, como resultado de recibir un enunciado pronunciado por una persona que habla;
obtener información indicativa de la identidad de la persona que habla;
interpretar la señal acústica al menos en parte determinando, utilizando la información indicativa de la identidad de la persona que habla y el reconocimiento automático de voz, si el enunciado pronunciado por la persona que habla incluye la al menos una palabra de activación designada; e
interactuar con la persona que habla en función, al menos en parte, de los resultados de la interpretación.

2. El sistema de la reivindicación 1 o cualquier otra reivindicación precedente, donde la interacción con la persona que habla comprende permitir a la persona que habla controlar la al menos una aplicación controlada por voz cuando el enunciado pronunciado por la persona que habla se ha determinado que incluye la al menos una palabra de activación designada.

3. El sistema de la reivindicación 1 o cualquier reivindicación precedente, donde la interpretación comprende además determinar, usando la información indicativa de la identidad de la persona que habla, si la persona que habla está autorizada para controlar la al menos una aplicación controlada por voz y
donde la interacción con la persona que habla comprende permitir que la persona que habla control la al menos una aplicación controlada por voz cuando se determina que la persona que habla está autorizada para controlar la al menos una aplicación controlada por voz y no permitir que la persona que habla controle la al menos una aplicación controlada por voz cuando se ha determinado que la persona que habla no está autorizada para controlar la al menos una aplicación controlada por voz.

4. El sistema de la reivindicación 1 o cualquier otra reivindicación precedente, donde la obtención de información indicativa de la identidad de la persona que habla comprende obtener la identidad de la persona que habla, al menos en parte, procesando la señal acústica, y
donde la obtención de la identidad de la persona que habla comprende:

obtener características del enunciado desde la señal acústica;
comparar las características del enunciado obtenido con las características del enunciado almacenado para cada una de las múltiples personas que hablan registradas en el sistema.

5. El sistema de la reivindicación 4 o cualquier otra reivindicación precedente, donde la determinación de si el enunciado pronunciado por la persona que habla incluye al menos una palabra de activación designada comprende:

identificar una o más palabras de activación asociadas con la identidad de la persona que habla; y
usar el reconocimiento por voz automatizado para determinar si el enunciado pronunciado por la persona que habla incluye una palabra de activación en las una o más palabras de activación, donde el reconocimiento por voz automatizado se realiza usando las una o más palabras de activación asociadas con la identidad de la persona que habla.

6. El sistema de la reivindicación 1 o cualquier otra reivindicación precedente, donde la obtención de información indicativa de la identidad de la persona que habla comprende determinar una posición de la persona que habla en un entorno.

7. El sistema de la reivindicación 6 o cualquier otra reivindicación precedente, donde el al menos un procesador de hardware de ordenador se configura además para realizar:

determinar, usando la posición de la persona que habla en el entorno, si la persona que habla está autorizada para controlar la al menos una aplicación controlada por voz, y
donde la interacción con la persona que habla comprende permitir que la persona que habla control la al menos una aplicación controlada por voz cuando se determina que la persona que habla está autorizada para controlar la al menos una aplicación controlada por voz y no permitir que la persona que habla controle la al menos una aplicación controlada por voz cuando se ha determinado que la persona que habla no está autorizada para controlar la al menos una aplicación controlada por voz.

8. El sistema de la reivindicación 6 o cualquier otra reivindicación precedente, donde la persona que habla está dentro de un vehículo y donde la determinación de la posición de la persona que habla se realiza, al menos en parte, en función de la información recopilada por al menos un sensor en el vehículo, donde el al menos un sensor es

diferente del al menos un micrófono.

9. El sistema de la reivindicación 1 o cualquier otra reivindicación precedente, donde el al menos un micrófono incluye una pluralidad de micrófonos y donde la determinación de la posición de la persona que habla se realiza usando señales acústicas generadas por la pluralidad de micrófonos.

10. El sistema de la reivindicación 1 o cualquier otra reivindicación precedente, donde el al menos un micrófono incluye un primer micrófono y un segundo micrófono, donde la señal acústica es generada por el primer micrófono y donde el al menos un procesador de hardware de ordenador está configurado además para realizar:

recibir una segunda señal acústica generada por el segundo micrófono al menos en parte como resultado de recibir, simultáneamente con el primer micrófono que recibe el enunciado pronunciado por la persona que habla, un segundo enunciado pronunciado por una segunda persona que habla;
obtener información indicativa de la identidad de la segunda persona que habla;
interpretar la segunda señal acústica al menos en parte determinando, utilizando la información indicativa de la identidad de la segunda persona que habla y el reconocimiento automático de voz, si el segundo enunciado pronunciado por la segunda persona que habla incluye una segunda palabra de activación designada para una segunda aplicación reconocida por voz; e
interactuar con la segunda persona que habla en función, al menos en parte, de los resultados de la interpretación.

11. El sistema de la reivindicación 1 o cualquier otra reivindicación precedente, donde la al menos una palabra de activación designada comprende una primera palabra de activación designada para una primera aplicación controlada por voz de la al menos una aplicación controlada por voz, y donde la primera palabra de activación designada es específica de la persona que habla, de modo que ninguna otra persona que habla puede usar la primera palabra de activación designada.

12. El sistema de la reivindicación 1 o cualquier otra reivindicación precedente, donde la obtención de información indicativa de la identidad de la persona que habla comprende obtener la identidad de la persona que habla; y donde la determinación de si el enunciado pronunciado por la persona que habla incluye la al menos una palabra de activación designada comprende:

acceder a una lista de palabras de activación asociadas con la identidad de la persona que habla; y
determinar si el enunciado incluye cualquier palabra de activación en la lista de palabras de activación asociadas con la identidad de la persona que habla.

13. El sistema de la reivindicación 1 o cualquier otra reivindicación precedente, donde la determinación, usando reconocimiento por voz automatizado, de si el enunciado pronunciado por la persona que habla incluye la al menos una palabra de activación designada comprende:
compensar la interferencia recibida por el al menos un micrófono utilizando la información asociada con la identidad de la persona que habla.

14. Un método para detectar al menos una palabra de activación designada para al menos una aplicación controlada por voz, comprendiendo el método:
usar al menos un procesador de hardware de ordenador para realizar:

recibir una señal acústica generada por al menos un micrófono, al menos en parte como resultado de recibir un enunciado pronunciado por una persona que habla;
obtener información indicativa de la identidad de la persona que habla;
interpretar la señal acústica al menos en parte determinando, utilizando la información indicativa de la identidad de la persona que habla y el reconocimiento automático de voz, si el enunciado pronunciado por la persona que habla incluye la al menos una palabra de activación designada; y
interactuar con la persona que habla en función, al menos en parte, de los resultados de la interpretación.

15. Al menos un medio de almacenamiento no transitorio legible por ordenador que almacena instrucciones ejecutables por el procesador que, cuando son ejecutadas por al menos un procesador de hardware de ordenador, hacen que el al menos un procesador de hardware de ordenador realice un método para detectar al menos uno palabra de activación designada para al menos una aplicación controlada por voz, comprendiendo el método:

recibir una señal acústica generada por al menos un micrófono, al menos en parte como resultado de recibir un enunciado pronunciado por una persona que habla;
obtener información indicativa de la identidad de la persona que habla;
interpretar la señal acústica al menos en parte determinando, utilizando la información indicativa de la identidad de la persona que habla y el reconocimiento automático de voz, si el enunciado pronunciado por la persona que habla incluye la al menos una palabra de activación designada; e
interactuar con la persona que habla en función, al menos en parte, de los resultados de la interpretación.

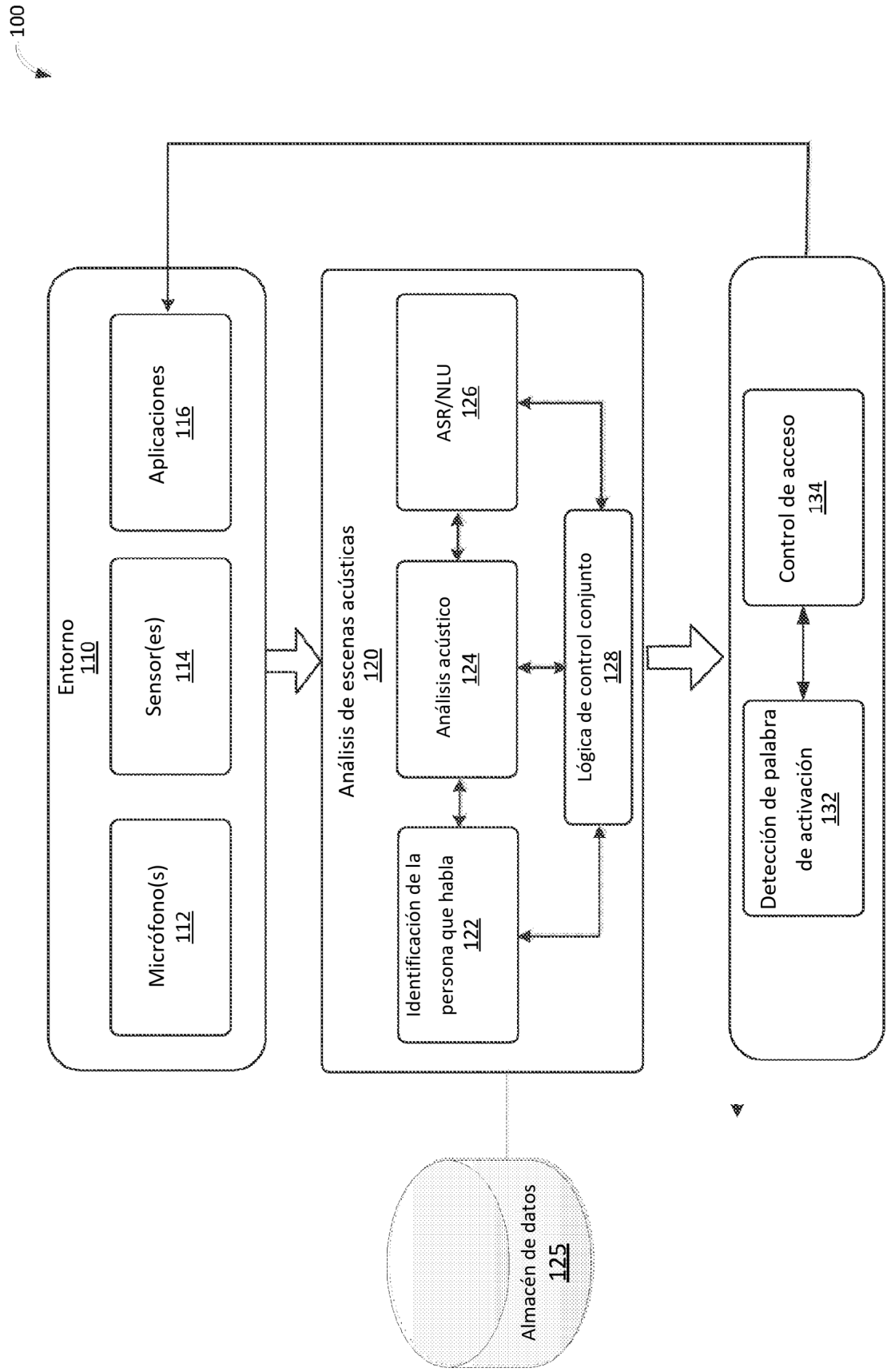
100
Entorno
110
Micrófono(s)
112
Sensor(es)
114
Aplicaciones
116
Análisis de escenas acústicas
120
Identificación de la persona que habla
122
Análisis acústico
124
ASR/NLU
126
Lógica de control conjunto
128
Detección de palabra de activación
132
Control de acceso
134
Almacén de datos
125

FIG. 1

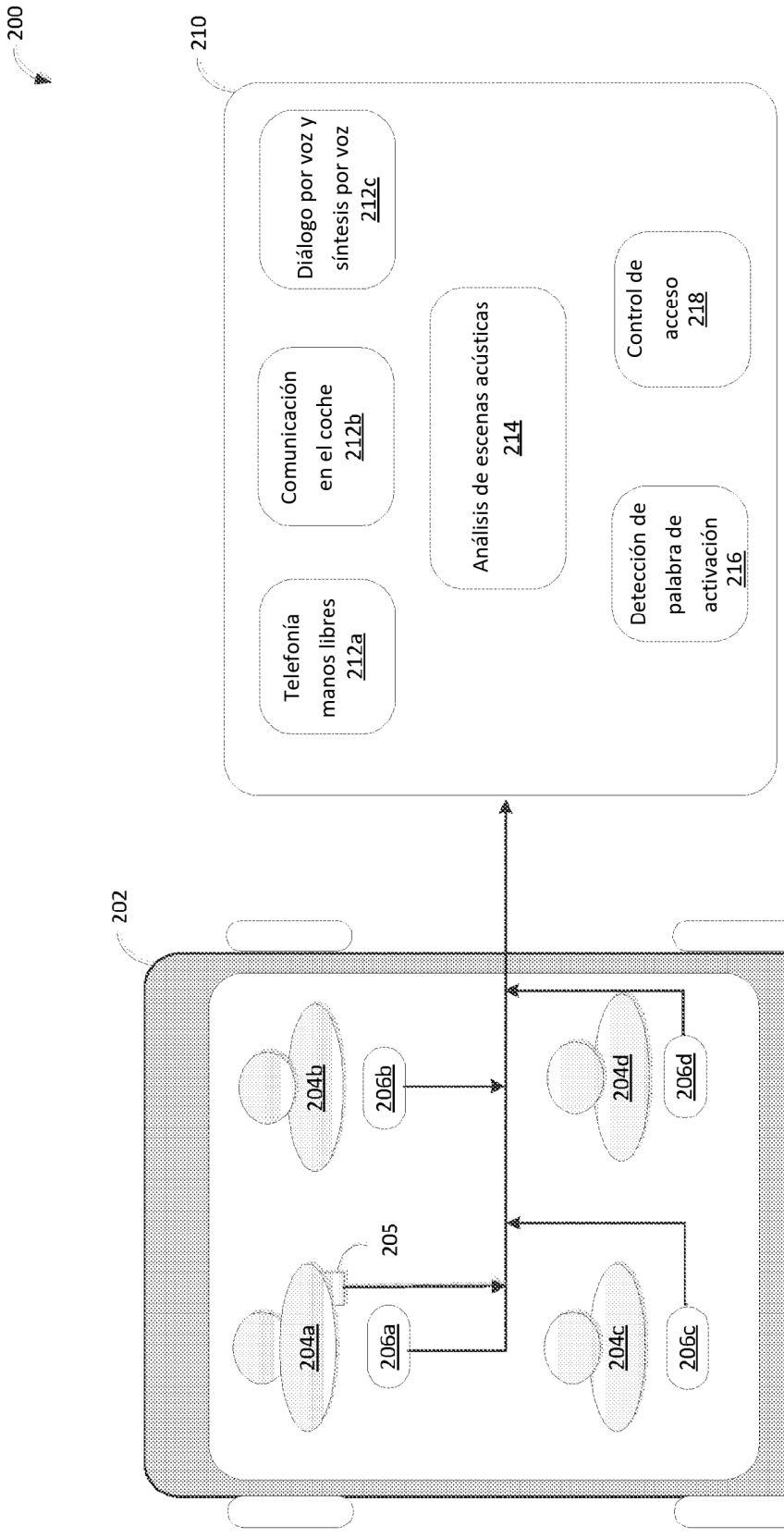
200
210
Diálogo por voz y síntesis por voz 212c
Comunicación en el coche 212b
Telefonía manos libres 212a
Análisis de escenas acústicas 214
Control de acceso 218
Detección de palabra de activación 216
202
204b
206b
204d
206d
205
204a
206a
204c
206c

FIG. 2

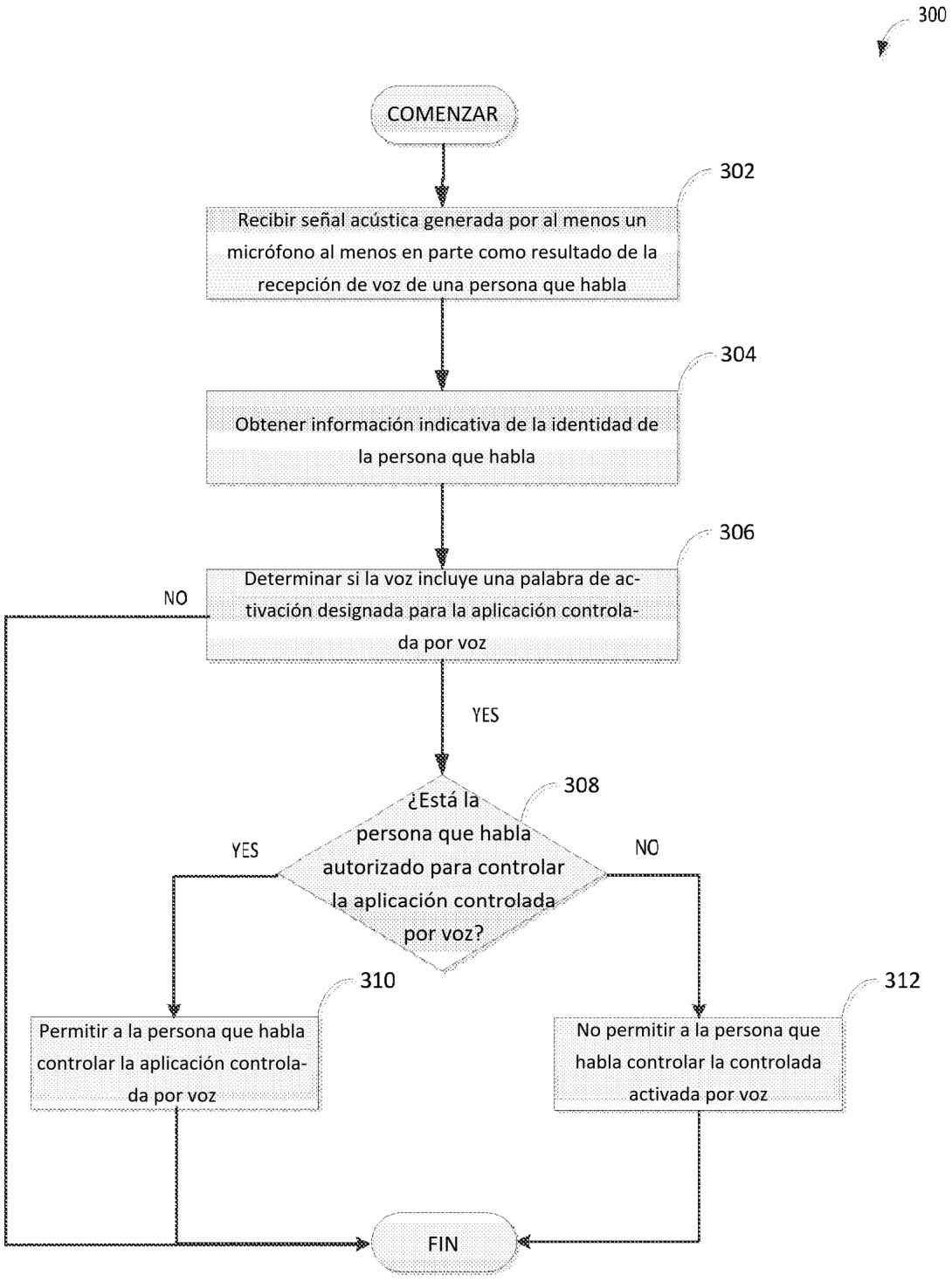
300
COMENZAR
302
Recibir señal acústica generada por al menos un micrófono al menos en parte como resultado de la recepción de voz de una persona que habla
304
Obtener información indicativa de la identidad de la persona que habla
306
Determinar si la voz incluye una palabra de activación designada para la aplicación controlada por voz
NO
YES
308
¿Está la persona que habla autorizado para controlar la aplicación controlada por voz?
YES
NO
310
Permitir a la persona que habla controlar la aplicación controlada por voz
312
No permitir a la persona que habla controlar la controlada activada por voz
FIN

FIG. 3

ID de la persona que habla	Posición	Aplicaciones autorizadas	Palabras de activación
Alice	Asiento del conductor, asiento del copiloto	Todas	"¡Oye!"
Bob	Asiento del conductor	Todas	"¡Despierta!"
Charlie	Asiento trasero; asiento del copiloto	Entretenimiento; telefonía	Entretenimiento - ("TV"); Teléfono - "Llama a alguien"
David	Asiento trasero; asiento del copiloto	Ninguna	Ninguna

FIG. 4

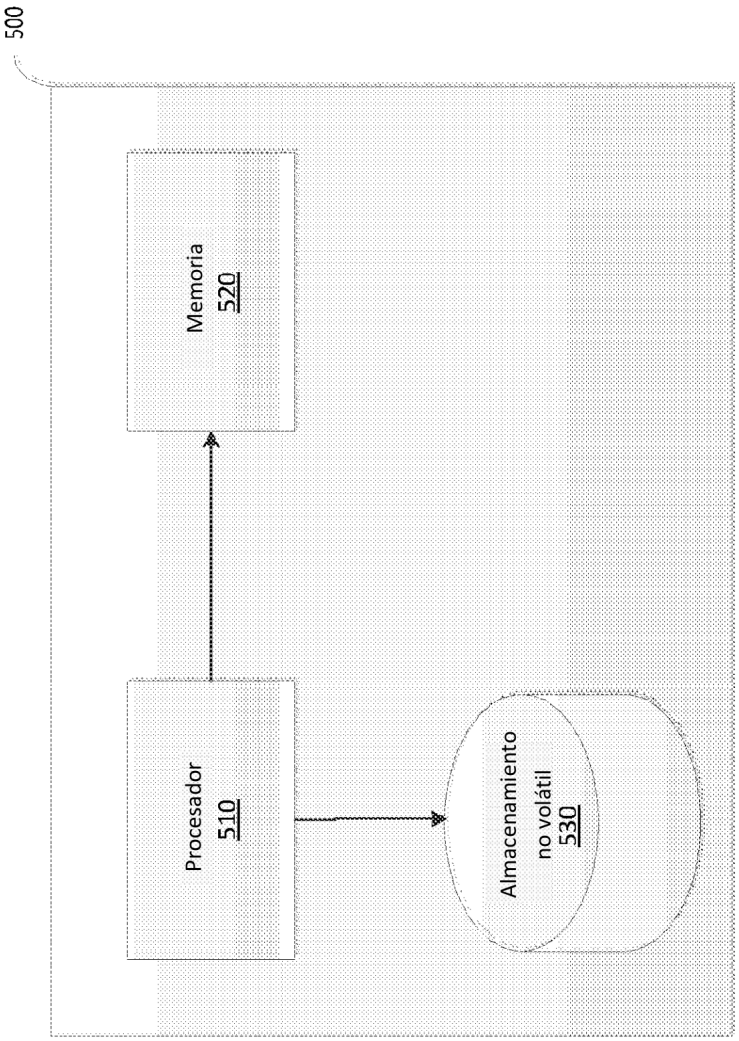
500
Memoria
520
Procesador
510
Almacenamiento
no volátil
530

FIG. 5