(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 751 484**

(51) Int. CI.:

G10L 15/22 (2006.01)
G06F 17/20 (2006.01)
G10L 25/54 (2013.01)
G06F 16/245 (2009.01)
G06F 16/432 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **07.05.2014 PCT/US2014/037080**

(87) Fecha y número de publicación internacional: **13.11.2014 WO14182771**

(96) Fecha de presentación y número de la solicitud europea: **07.05.2014 E 14795114 (9)**

(97) Fecha y número de publicación de la concesión europea: **28.08.2019 EP 2994908**

(54) Título: **Interfaz de entrada de voz incremental con retroalimentación en tiempo real**

(30) Prioridad:

07.05.2013 US 201361820267 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
31.03.2020

(73) Titular/es:

VEVEO, INC. (100.0%)
2160 Gold Street
San Jose, CA 95002, US

(72) Inventor/es:

ARAVAMUDAN, MURALI;
WELLING, GIRISH;
GILL, DAREN;
ARDHANARI, SANKAR;
BARVE, RAKESH y
VENKATARAMAN, SASHIKUMAR

(74) Agente/Representante:

PONS ARIÑO, Ángel

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Interfaz de entrada de voz incremental con retroalimentación en tiempo real

CAMPO DE LA DESCRIPCIÓN

La presente descripción se refiere a sistemas y procedimientos para ayudar a un usuario a recuperar información aplicando una entrada incremental a una interfaz conversacional y, más específicamente, relacionada con técnicas para proporcionar una retroalimentación interactiva a un usuario durante la entrada incremental a una interfaz conversacional.

ANTECEDENTES DE LA DESCRIPCIÓN

El descubrimiento de contenido basado en la entrada de voz se encuentra en una etapa similar de evolución en comparación con las interfaces de entrada basadas en texto hace casi una década. Un usuario expresa su intención diciendo una oración completamente formada y a continuación espera una respuesta. La respuesta puede hacer que el usuario conteste con otra oración completa. De manera análoga a este modelo de uso, hace casi una década, el usuario expresaba toda su intención en forma de palabras clave completamente formadas, y a continuación enviaba la consulta de búsqueda introducida completamente. La búsqueda incremental basada en texto cambió este paradigma operativo. La búsqueda incremental basada en texto se describe más adelante en la patente de EE. UU. n.º 7.895.218 denominada *"Procedimiento y sistema para realizar búsquedas de contenido de televisión utilizando una entrada de texto reducida"* y presentada el 24 de mayo de 2005. En la búsqueda incremental basada en texto, los resultados de búsqueda aparecen cuando el usuario escribe palabras clave (o incluso simples prefijos que corresponden a palabras clave). Los usuarios ahora dan por sentado la facilidad de uso de una interfaz de búsqueda incremental basada en texto.

Los sistemas de descubrimiento de contenido basados en voz se están volviendo lo suficientemente fiables y útiles como para incorporarse a la vida cotidiana de los usuarios. Si bien los usuarios se han visto condicionados por la facilidad de uso de la búsqueda incremental basada en texto, el descubrimiento de contenido basado en voz también está marcando el comienzo de un cambio lento en la expresión de la intención. Por ejemplo, el descubrimiento de contenido basado en voz ofrece la capacidad de verbalizar directamente las mentes de los usuarios, en lugar de traducir el pensamiento a una cadena de palabras clave. Si bien las interfaces de voz basadas en el lenguaje natural se encuentran principalmente en el entorno móvil y de la televisión, el entorno de escritorio también está asistiendo a la aparición de interfaces de lenguaje natural, como la Graph Search de Facebook, donde el usuario escribe las consultas en lenguaje natural.

La patente de EE. UU. 2006/0206454 A1 describe un sistema donde se recibe la entrada de búsqueda desde un campo de búsqueda de una aplicación de navegador web o de las palabras pronunciadas por un usuario y se convierten a texto usando un software de reconocimiento de voz. Según las características de la entrada de búsqueda, se determina si se debe enviar automáticamente una consulta a un motor de búsqueda. La patente de EE. UU. 2011/0145224 A1 describe un sistema para recibir una consulta de búsqueda por voz del usuario y reconocer e identificar incrementalmente los términos de búsqueda. Después de que la consulta se haya reconocido de forma incremental, el sistema utilizará los términos de búsqueda para recuperar una parte de los resultados de búsqueda en función de los términos de búsqueda utilizables identificados. La patente de EE. UU. 2010/0153112 A1 describe un procedimiento de búsqueda donde la voz del usuario se traduce en una consulta textual y se envía a un motor de búsqueda. Los resultados de la búsqueda se presentan al usuario. A medida que el usuario continúa hablando, la consulta de voz se refina en función de la posterior conversación del usuario. La consulta de voz refinada se convierte en una consulta textual que nuevamente se envía al motor de búsqueda.

RESUMEN

Las siguientes instancias de la palabra "realización(es)", si se refieren a combinaciones de características diferentes de las definidas por las reivindicaciones independientes, se refieren en cualquier caso a ejemplos que se presentaron originalmente pero que no representan las realizaciones de la invención actualmente reivindicada; estos ejemplos se muestran, no obstante, solo con fines ilustrativos. La presente descripción incluye procedimientos y sistemas para seleccionar y presentar elementos de contenido basados en la entrada del usuario. La presente descripción presenta una interfaz de entrada incremental para la recuperación de información. Los presentes sistemas y procedimientos proporcionan retroalimentación sobre la interpretación del sistema de la entrada del usuario y devuelven respuestas basándose en esa interpretación.

Según un aspecto, un procedimiento implementado por ordenador para seleccionar y presentar los elementos de contenido basados en la entrada del usuario comprende proporcionar acceso a un conjunto de elementos de contenido

asociados a metadatos que describe el elemento de contenido correspondiente, recibir una primera entrada con la que el usuario tiene la intención de identificar al menos un elemento de contenido deseado, determinar que al menos una parte de la primera entrada tiene importancia en un grado que excede un valor umbral, proporcionar retroalimentación al usuario que identifica la parte de la entrada y recibir una segunda entrada del usuario posterior a la primera, deducir si el usuario tenía la intención de modificar la primera entrada con la segunda o de complementar la primera entrada con la segunda, con la condición de la que se deduce que el usuario tenía la intención de modificar la primera entrada con la segunda, determinar una consulta alternativa que combina la primera entrada con la segunda, con la condición de la que se deduce que el usuario tenía la intención de complementar la primera entrada con la segunda, determinar una consulta alternativa que combina la primera entrada y la segunda, seleccionar un subconjunto de elementos de contenido del conjunto de elementos de contenido basado en la comparación de la consulta alternativa y los metadatos asociados con el subconjunto de elementos de contenido y presentar el subconjunto de elementos de contenido al usuario, donde la deducción de si el usuario tenía la intención de modificar la primera entrada con la segunda o de complementar la primera entrada con la segunda puede incluir la determinación de un grado de similitud entre la primera entrada y la segunda, con la condición de que ese grado de similitud esté por encima de un umbral, deducir que el usuario tenía la intención de modificar la primera entrada y, con la condición de que el grado de similitud esté por debajo de un umbral, deducir que el usuario tenía la intención de complementar la primera entrada.

El aspecto anterior también puede proporcionar un sistema para seleccionar y presentar elementos de contenido basados en la entrada del usuario, comprendiendo el sistema instrucciones legibles por ordenador codificadas en un medio legible por ordenador no transitorio, las instrucciones legibles por ordenador que hacen que el sistema informático esté configurado para proporcionar acceso a un conjunto de elementos de contenido, estando dichos elementos de contenido asociados a metadatos que describen un elemento de contenido correspondiente, recibir una primera entrada con la que el usuario tiene la intención de identificar al menos un elemento de contenido deseado, determinar que al menos una parte de la primera entrada tiene un grado de importancia que excede un valor umbral, proporcionar retroalimentación al usuario que identifica la parte de la entrada y recibir una segunda entrada del usuario posterior a la primera, deducir si el usuario tenía la intención de modificar la primera entrada con la segunda o complementar la primera entrada con la segunda, con la condición de la que se deduce que el usuario tenía la intención de modificar la primera entrada con la segunda, determinar una consulta alternativa que combina la primera entrada y la segunda, con la condición de la que se deduce que el usuario tenía la intención de complementar la primera entrada con la segunda, determinar una consulta alternativa que combina la primera entrada con la segunda, seleccionar un subconjunto de elementos de contenido del conjunto de elementos de contenido basado en la comparación de la consulta alternativa y los metadatos asociados con el subconjunto de elementos de contenido y presentar el subconjunto de elementos de contenido al usuario, donde la deducción de si el usuario tenía la intención de modificar la primera entrada con la segunda o complementar la primera entrada con la segunda puede incluir la determinación de un grado de similitud entre la primera entrada y la segunda, con la condición de que ese grado de similitud esté por encima de un umbral, deducir que el usuario tenía la intención de modificar la primera entrada, y con la condición de que el grado de similitud esté por debajo de un umbral, deducir que el usuario tenía la intención de complementar la primera entrada.

Las realizaciones descritas en la presente invención pueden incluir aspectos adicionales. Por ejemplo, la determinación de que al menos la parte de la primera entrada tiene un grado de importancia que excede el valor umbral incluye la identificación de uno o más límites de frase en la entrada incremental, y la identificación de uno o más límites de la frase se basa al menos en parte o al menos en uno de los siguientes: (a) una disfluencia identificada en la primera entrada del usuario, (b) reglas gramaticales aplicadas a la primera entrada, (c) el grado de importancia de la parte de la primera entrada, (d) al menos una interacción conversacional previa con el usuario, y (e) una firma de preferencia del usuario. La firma de preferencia del usuario puede describir las preferencias del usuario para al menos uno de (i) los elementos de contenido en particular y (ii) los metadatos en particular asociados con los elementos de contenido, donde la parte de la primera entrada se identifica en función de la firma de preferencia del usuario. La disfluencia puede incluir una pausa en la entrada de voz, un relleno de tiempo auditivo en la entrada de voz y/o una pausa en la escritura. La selección del subconjunto de elementos de contenido puede basarse además en una disfluencia identificada en la primera entrada, y también en interacciones conversacionales previas que se determina que están relacionadas con la primera y la segunda entrada. En la retroalimentación proporcionada se puede incluir la solicitud de una aclaración sobre la parte identificada de la entrada, sugerir la finalización de la primera entrada recibida y/o la repetición de la parte de la entrada al usuario, para notificar al usuario que parte de la entrada puede haberse reconocido de manera incorrecta. La solicitud de aclaración sobre la parte identificada de la entrada puede basarse, al menos en parte, en una determinación de que se produce una disfluencia después de que el usuario haya proporcionado parte de la entrada. La sugerencia de completar la primera entrada recibida puede basarse, al menos en parte, en una determinación de que la disfluencia ocurre antes de que se espere que el usuario proporcione la parte de la entrada. La retroalimentación proporcionada al usuario puede elegirse en función de la duración de una disfluencia identificada en la primera entrada, un grado de confianza en el correcto reconocimiento de voz a texto de la parte de la entrada, un recuento de las ambigüedades detectadas en la primera entrada, un recuento las de

correcciones de errores necesarias para identificar la parte de la entrada, un recuento de los nodos en la estructura de datos gráficos, donde dicho recuento de los nodos en la estructura de datos gráficos mide la ruta entre un primer nodo que representa un elemento de interés de una interacción conversacional previa y un segundo nodo que representa la parte de la entrada, y/o un grado de relación de la parte de la entrada con interacciones conversacionales previas con el usuario. La presentación del subconjunto de elementos de contenido puede incluir la presentación del subconjunto de elementos de contenido antes de recibir una entrada completa del usuario, al determinar una fuerte coincidencia de reconocimiento para la primera entrada y al determinar que un grado de respuesta del subconjunto seleccionado de elementos de contenido estaría por encima de un umbral.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Para una comprensión más completa de varias realizaciones de los presentes sistemas y procedimientos, a continuación se hace referencia a las siguientes descripciones tomadas en relación con los dibujos adjuntos, donde los números de referencia similares se refieren a elementos similares:

La figura 1 ilustra los componentes de entrada y salida de un sistema de ejemplo para una entrada incremental, según algunas realizaciones.

La figura 2 ilustra un sistema de ejemplo para una interfaz de entrada de voz incremental, según algunas realizaciones.

La figura 3 representa un flujo de datos de ejemplo de los presentes sistemas, según algunas realizaciones.

Las figuras 4-5 ilustran ejemplos de interacciones conversacionales entre el presente sistema y un usuario, según algunas realizaciones.

Las figuras 6-8 ilustran ejemplos de muestras de formas de onda para la introducción de datos del usuario “Quién actuaba en la película *El mañana nunca muere,”* según algunas realizaciones.

Las figuras 9-10 ilustran formas de onda de ejemplo que muestran el uso de la falta de pausas para detectar el recorte en la entrada de voz, según algunas realizaciones.

DESCRIPCIÓN DETALLADA DE LAS REALIZACIONES PREFERIDAS

Vista general

La presente descripción proporciona una interfaz de entrada incremental para la recuperación de información, donde los sistemas y procedimientos presentes proporcionan retroalimentación en tiempo real de la interpretación del sistema de la entrada del usuario y la devolución de respuestas en función de esa interpretación. Algunas realizaciones incluyen una interfaz de entrada incremental basada en voz, donde los presentes sistemas y procedimientos proporcionan retroalimentación en tiempo real sobre la entrada del usuario mientras este habla. Los presentes procedimientos y sistemas permiten una experiencia de usuario similar a las interacciones humanas donde un oyente responde a una consulta inmediatamente o incluso antes de que el usuario finalice una pregunta.

Además de centrarse en la experiencia natural de las interacciones humanas, los presentes sistemas y procedimientos abordan las muchas deficiencias que afectan a los sistemas tradicionales basados en la conversación, por ejemplo, donde abundan los errores en el reconocimiento de voz. La retroalimentación proporcionada por los presentes sistemas permite al usuario saber en tiempo real si se produce un error y tener así la oportunidad de corregirlo. El usuario no tiene que comenzar a hablar de nuevo, o recurrir a un modo de interacción de entrada de texto, para editar la larga cadena de texto de la entrada de voz. Es sabido que el uso del modo de entrada de texto es engorroso y peligroso en un entorno móvil, dada la exclusiva atención cognitiva, motora y sensorial que exige la edición de texto. Los requisitos de ejemplo de un modo de entrada de texto incluyen colocar o navegar al punto de inserción de texto en medio de una cadena para eliminar una frase o palabra incorrecta y a continuación escribir la frase o palabra correcta. Los presentes procedimientos y sistemas proporcionan retroalimentación en tiempo real sobre la entrada del usuario en los límites de la frase, y también devuelven respuestas completas o una sinopsis de respuestas a la entrada del usuario.

Algunas realizaciones proporcionan las respuestas en forma de audio y/o visual que coinciden con la forma donde el usuario está utilizando el presente sistema. Al igual que en el caso de las interacciones humanas, donde las pausas en las conversaciones son señales importantes para calibrar el nivel de comprensión, la confianza y/o la falta de confianza en el contenido que se expresa, los presentes sistemas y procedimientos hacen un uso completo de las disfluencias en la conversación, dentro y entre las oraciones, interpretando adecuadamente dichas disfluencias para sincronizar la retroalimentación auditiva/visual y las respuestas al usuario. Las disfluencias de ejemplo incluyen pausas

en la conversación; rellenos auditivos en la conversación como “um”, “hmm”, “esto”, “uh”, “eh”, “bueno” o incluso “como”; o pausas en la introducción de datos con el teclado, como cuando un usuario se detiene al escribir.

Descubrimiento del problema y ventajas de la solución

Los solicitantes han descubierto que desde la perspectiva del procedimiento de entrada, la expresión de la intención como una oración completamente formada en lugar de una cadena de palabras clave, requiere que las interfaces de voz para el descubrimiento de contenido no sean como motores de ejecución de “comandos de voz”. Necesitan dar sentido a oraciones completas para generar respuestas.

Los solicitantes también han descubierto que otro desafío al que se enfrentan los sistemas de descubrimiento de contenido basado en voz es la corrección de errores que se deslizan en la entrada del usuario debido a los errores de reconocimiento de voz. Pronunciar una oración completa y a continuación tener que corregir una frase en medio de una oración convertida de voz a texto no es fácil, particularmente en un entorno móvil donde el usuario puede estar conduciendo y tener las manos y la vista completamente ocupados. Incluso ignorando lo engorroso de la edición de texto, la atención cognitiva, motora y sensorial exclusiva que exige la operación de edición descarta la edición de texto como una opción para la corrección en entornos móviles, como cuando el usuario está conduciendo.

Además, los solicitantes han descubierto que un desafío más sutil que abordar en una interfaz de descubrimiento de contenido basado en la entrada de voz es igualar la facilidad de uso que ofrece de manera natural la búsqueda incremental basada en texto. En la búsqueda incremental basada en texto, a medida que el usuario empieza a escribir, los resultados proporcionan información visual instantánea del sistema que converge en los resultados. Es decir, cuando el usuario escribe la frase “programador de Java” en una entrada de búsqueda de texto, el usuario comienza a ver resultados de búsqueda de programadores de JAVA, por ejemplo, en un sistema de recuperación de información para selección de personal. La posterior adición de restricciones como “Boston” o “desarrollo de sistemas integrados” podría ser una elección dinámica que el usuario puede hacer en función de los resultados que aparecen a medida que escribe. La naturaleza instantánea de la respuesta a la búsqueda tiene dos objetivos clave: (1) la respuesta ofrece retroalimentación instantánea al usuario y (2) la respuesta incluso ayuda al usuario a adaptar las preguntas posteriores en función de la respuesta, todo ello mientras escribe en un cuadro de búsqueda. No es suficiente con que el rango de precisión del reconocimiento de voz mejore para igualar la facilidad de uso de la búsqueda incremental basada en texto. Incluso si el reconocimiento de voz se aproxima al 100 % de precisión, la capacidad de los presentes sistemas y procedimientos de modificar dinámicamente la intención original en función de las respuestas a la entrada parcial es un factor útil de experiencia del usuario para calibrar la capacidad de respuesta y la inteligencia del sistema. Las conversaciones que muchos usuarios clasificarían como interesantes aún pueden contener interrupciones e interjecciones mutuas. Por ejemplo, estas interrupciones e interjecciones forman la esencia de la riqueza de la conversación, incluso si la conversación se mantiene centrada en el tema principal a tratar.

Las realizaciones de los presentes sistemas y procedimientos abordan los desafíos descritos anteriormente de múltiples maneras. Un procedimiento sirve para dar al usuario la confianza de que el reconocimiento de voz ha funcionado tanto de forma visual como auditiva. Esta confirmación puede ser por una respuesta auditiva y/o visual (*p. ej*., a través del tablero de instrumentos de un automóvil o de la pantalla del teléfono móvil). En forma auditiva, esta confirmación puede incluir la repetición de frases reconocidas (no todas las palabras del usuario) o proporcionar una sinopsis de la respuesta, mientras el usuario habla (*p. ej.,* los presentes sistemas responden con “50 coincidencias de programadores de Java” como resultado de búsqueda). El presente sistema también puede generar resultados completos de forma visual o auditiva cuando el usuario hace una pausa; la duración de una pausa se usa como una métrica para decidir qué resultados presentar.

En resumen, la presente descripción presenta una interfaz de descubrimiento de contenido basada en la entrada de voz donde la retroalimentación y las respuestas se presentan cuando el usuario expresa su intención, donde las respuestas pueden incluir retroalimentación sobre la entrada del usuario y los resultados de la intención expresada por el usuario hasta ese momento.

Sistema de entrada incremental

Las realizaciones preferidas de la presente invención y sus ventajas pueden entenderse consultando las figuras 1-10.

La figura 1 ilustra los componentes de entrada y salida de un sistema de ejemplo 103 para una entrada incremental, según algunas realizaciones. El sistema 103 incluye las entradas de micrófono 101 y la interfaz visual 102. El sistema 103 también incluye las salidas de auriculares/altavoz 104 y la pantalla 105

Algunas realizaciones permiten al usuario expresar su intención al hablar o escribir. El micrófono 101 y la interfaz visual 102, como un cuadro de texto o equivalente, pueden habilitar ambas formas de entrada. Otras realizaciones

incluyen ambas formas de entrada, o solo una. De forma similar, algunas realizaciones presentan respuestas visuales y auditivas, respectivamente, en la pantalla 105 y los auriculares/altavoz 104. Entre los dispositivos que admiten diferentes combinaciones de estas entradas y salidas se encuentran las tabletas (p. ej., iPad) que tienen ambas formas de entrada y salida, teléfonos móviles que tienen ambas formas de entrada y salida (aunque en el uso real, el usuario solo puede usar entradas y salidas de audio, por ejemplo, mientras conduce), las pantallas del tablero de instrumentos del automóvil que tienen entrada de audio y salida tanto de imagen como de audio, los ordenadores de sobremesa/tabletas que tienen ambas formas de entrada y salida (aunque el usuario solo puede usar expresamente los medios de entrada y salida visual).

El último caso de uso de ordenadores de sobremesa/tabletas que admiten entrada y salida de audio e imagen puede parecer inicialmente un simple caso de uso de búsqueda incremental basada en texto. Sin embargo, los presentes sistemas y procedimientos todavía pueden usar la entrada de lenguaje natural en combinación con la capacidad de detectar términos de importancia en los límites de la frase, para facilitar la edición rápida. Por ejemplo, las interfaces visuales de los dispositivos táctiles dificultan la interacción a la hora de colocar un punto de inserción para eliminar o seleccionar. Incluso en los ordenadores de sobremesa, donde las operaciones de edición son más fáciles, algunas realizaciones hacen que la interfaz de entrada de texto mediante lenguaje natural sea una extensión sencilla y natural del familiar cuadro de texto de búsqueda. Por ejemplo, los presentes sistemas interpretan las disfluencias, p. ej., las pausas detectadas entre las oraciones mientras el usuario escribe que representan los límites de las oraciones. Este uso de las disfluencias elimina la necesidad de que el usuario introduzca delimitadores explícitos, como la puntuación. Por consiguiente, en algunas realizaciones, el sistema 103 amplía la conocida interfaz visual basada en texto 102 como un medio para introducir la intención en la presente interfaz de lenguaje natural. El presente sistema utiliza las disfluencias detectadas entre las palabras y oraciones para interpretar los límites de las oraciones y la conversación. Los presentes sistemas se convierten en interfaces familiares y fáciles de adoptar, tanto para usuarios que desean utilizar interfaces de lenguaje natural como para usuarios que ya se sienten cómodos con el refinamiento de búsqueda basado en palabras clave. La determinación implícita de los límites de las oraciones y la conversación basados en las disfluencias hace posible esta familiaridad. Otras realizaciones usan modos de texto y voz híbridos o combinados para permitir que un usuario exprese su intención, si lo desea.

La figura 2 ilustra un sistema de ejemplo 213 para una interfaz de entrada de voz incremental, según algunas realizaciones. El sistema 213 incluye el reconocedor de voz 201, la interfaz de entrada de texto 203, la interfaz de representación de retroalimentación y respuesta 204 y el convertidor de texto a voz 206.

El sistema 213 alimenta el flujo de entrada de voz 200 desde el usuario al reconocedor de voz 201. El reconocedor de voz 201 emite texto reconocido en tiempo real mientras el usuario habla. En algunas realizaciones, la pantalla 205 presenta el texto emitido directamente como texto ya introducido en la interfaz de entrada de texto 203. En otras realizaciones, el sistema 213 envía el texto reconocido con errores al motor de conversación 202 (a través de la interfaz de entrada/salida 212). El motor 202 de conversación interpreta la salida reconocida, incluidos los errores incrustados, en el contexto de la conversación para realizar una corrección de errores más inteligente de la entrada del usuario. La interfaz de entrada de texto 203 muestra a continuación la salida reconocida. Al usar el contexto de la conversación, algunas realizaciones usan variables de estado previamente almacenadas sobre las interacciones presentes en el contexto de la conversación y/o conversaciones previas (p. ej., entidades, intenciones y/o resultados de interacciones conversacionales) para mejorar la predicción de qué intención intentaba comunicar el usuario. Por ejemplo, si un usuario dice una frase que contiene una palabra que el reconocedor de voz 201 podría interpretar como “java” o “jabba”, el sistema 213 puede deducir que el usuario estaba preguntando sobre “JAVA”, el lenguaje de programación, basándose en las interacciones conversacionales de anteriores solicitudes del usuario para obtener una lista de programadores disponibles en un área geográfica designada. En contraste, el sistema 213 deduce que el usuario no estaba solicitando elementos de contenido relacionados con “Jabba”, el nombre de un personaje ficticio de las películas de *Star Wars* (es decir, “Jabba el Hutt”). La corrección de errores descrita anteriormente y la determinación de las variables de estado correspondientes se describen en la patente de los EE. UU. n.º 2014/0108453 A1 denominada *“Procedimiento para la gestión adaptativa del estado de la conversación con operadores de filtrado aplicados dinámicamente como parte de una interfaz conversacional”,* presentada el 13 de marzo de 2013.

En algunas realizaciones, el sistema 213 resalta las frases como elementos interactivos que son fácilmente editables, basándose en el procesamiento actual de la entrada de voz 200 y el texto correspondiente en tiempo real del motor de conversación 202 antes de visualizarse en la interfaz de entrada de texto 203. El presente resaltado permite al sistema 213 manejar los errores de reconocimiento que no fueron descubiertos por el motor de conversación 202. En algunas realizaciones, el sistema 213 utiliza reglas de gramática para identificar los elementos interactivos apropiados a resaltar. Por ejemplo, el sistema 213 resalta los sustantivos o sujetos de una oración, según lo identificado por un motor de reglas gramaticales (no mostrado) o por otros procedimientos conocidos. En otras realizaciones, la interfaz de usuario 203 muestra la cadena de voz reconocida sin alterar, y la interfaz de representación de retroalimentación y respuesta 204 presenta información sobre la entrada de texto 208 desde el motor de conversación 203 por separado. Esta representación por separado permite al usuario editar fácilmente la cadena original convertida en texto.

La salida de ejemplo del motor de conversación 202 incluye la respuesta de voz 210 y la respuesta visual 209 a mostrar. El componente de voz 211 de la respuesta de voz 210 proporciona retroalimentación al usuario que identifica una parte de la entrada que el sistema 213 ha determinado que es importante. La retroalimentación de ejemplo incluye (1) solicitar aclaraciones sobre la parte de la entrada, (2) repetir la parte de la entrada, (3) sugerir que se complete la parte de la entrada del usuario, (4) proporcionar un breve resumen (p. ej., una "sinopsis") de los elementos de contenido solicitados en respuesta a la entrada de voz 200, y/o (5) proporcionar el subconjunto completo de elementos de contenido solicitados encontrados en respuesta a la entrada de voz 200. Un ejemplo de solicitud de aclaración sería que el sistema 213 preguntara "¿quiso decir programadores de "Java" o "Jabba el Hutt"?" Un ejemplo de repetición de la parte de la entrada sería que el sistema 213 repitiera una parte de la entrada reconocida incorrectamente, por ejemplo, "área de Bolton", cuando la entrada de voz 200 incluía "área de Boston". Un ejemplo de sugerencia de finalización sería que, en respuesta a la entrada de voz 200 "en qué película actuó Jessica Chastain dirigida por Terrence <pausa>", el sistema 213 respondiera "Terrence Malick"). Un ejemplo de sinopsis de los elementos de contenido solicitados sería que el sistema 213 respondiera "más de 100 programadores de Java en el área de Boston". Un ejemplo donde se proporciona todo el subconjunto de elementos de contenido solicitados sería que el sistema 213 enumerara los programadores de Java encontrados. El convertidor de texto a voz 206 convierte el componente de voz 210 del motor de conversación 202 en la salida de voz 211.

En algunas realizaciones, el convertidor de texto a voz 206 también alimenta la salida de voz 207 al reconocedor de voz 201, por ejemplo, para representarla en la salida de audio, de modo que el reconocedor de voz 201 pueda filtrar la salida del sistema a partir de la voz del usuario. Este filtrado asume que el sistema 213 está diseñado para evitar el tradicional timbre de bucle de retroalimentación de audio. Si el sistema 213 alimenta la salida de audio directamente a la toma de salida de los auriculares, esta retroalimentación de audio no es necesaria, ya que el usuario escucha la salida del sistema directamente, sin que el micrófono la detecte. En algunas realizaciones, la salida de audio del sistema tiene menos prioridad y se apaga cuando el sistema 213 detecta que el usuario está hablando, para no interrumpirlo. En otras realizaciones, el sistema puede no apagar la salida de audio, por ejemplo en los casos donde la salida de audio es la de los auriculares. Esta situación simula conversaciones de la vida real donde una persona interrumpe a otra. Por ejemplo, el sistema 213 "interrumpe" al usuario al determinar un grado de confianza que al presentar los resultados solicitados al usuario (es decir, al presentar el subconjunto solicitado de elementos de contenido) eliminaría la necesidad de que el usuario termine de proporcionar el resto de la entrada de voz 200. Esta funcionalidad puede beneficiarse de un diseño cuidadoso desde la perspectiva de la interfaz, para evitar que el sistema 213 le parezca "grosero" al usuario.

La figura 3 representa un flujo de datos 300 de ejemplo de los presentes sistemas, según algunas realizaciones. El presente sistema recibe la entrada de texto 302. En algunas realizaciones, la entrada de texto 302 se recibe como entrada de texto escrita directamente por el usuario. En otras realizaciones, la entrada de texto 302 se determina usando la conversión de voz a texto (paso 303) de la voz de usuario 301. El presente sistema marca en el tiempo la entrada de texto 302 con la información de las pausas dentro y entre las oraciones (paso 304). Algunas realizaciones utilizan el contexto conversacional para interpretar la información de las pausas (paso 305). El presente sistema utiliza la interpretación resultante para generar la retroalimentación o los resultados en respuesta al usuario (paso 306). A continuación se describen las técnicas para interpretar las pausas en el discurso, en relación con las figuras 6-10. Las técnicas para interpretar las pausas en el discurso se exponen con más detalle en la patente de los EE. UU. n.º 2014/0039895 A1 denominada *Procedimiento para utilizar las pausas detectadas en la entrada de voz para ayudar a interpretar la entrada durante la interacción conversacional para la recuperación de información,* presentada el 13 de marzo de 2013. Esas técnicas pueden utilizarse junto con las técnicas descritas en este documento. En algunas realizaciones, la respuesta generada incluye un componente visual 307 y un componente de voz. El presente sistema reproduce el componente de voz 308 al usuario y también utiliza el componente de voz como retroalimentación auditiva 309. Por ejemplo, el presente sistema cancela la señal reproducida en el caso de que una entrada de micrófono pueda detectar la retroalimentación auditiva 309.

La figura 4 ilustra un ejemplo de interacción conversacional 400 entre el presente sistema y un usuario, según algunas realizaciones. El usuario dice "muéstrame programadores de Java en el área de Boston" (intercambio 401). En cualquier momento, el presente sistema permite que el usuario pueda proporcionar disfluencias tales como hacer una pausa en cualquier punto para recibir retroalimentación del presente sistema, si el usuario duda de que se le haya entendido. Por ejemplo, si el usuario hace una pausa después de decir "muéstrame programadores" o "muéstrame programadores de java", el presente sistema determina que una parte de la entrada recibida hasta el momento tiene un grado de importancia que excede un valor umbral. El presente sistema proporciona retroalimentación y/o indicaciones sobre la parte de la entrada que se considera que tiene el grado de importancia, como "programadores de Java" o "área de Boston". Por ejemplo, el presente sistema proporciona retroalimentación sobre partes de la entrada acotadas por límites de la frase, o generalmente sobre cualquier parte de la entrada considerada importante. Algunas realizaciones pueden determinar el grado de importancia en función de los metadatos relacionados almacenados en una estructura de datos teóricos gráficos de nodos y entidades, descritos con más detalle a continuación. En algunas

realizaciones, el grado de importancia identifica las partes de la entrada que podrían beneficiarse de la aclaración, desambiguación o confirmación por parte del presente sistema. El presente sistema proporciona retroalimentación al usuario identificando la parte de la entrada considerada importante (es decir, el presente sistema proporciona retroalimentación o confirmación sobre las frases o términos importantes dichos hasta ese momento).

Al proporcionar la confirmación, el presente sistema funciona casi como un asistente personal que copia al dictado y repite ciertas frases o palabras pronunciadas, de modo que la persona que dicta es consciente de que la transcripción del texto del asistente personal se está realizando correctamente. Por lo tanto, en algunas realizaciones, el presente sistema comienza repitiendo la parte de la entrada considerada importante (es decir, repitiendo las frases dichas por el usuario). En realizaciones adicionales, el presente sistema utiliza el índice de éxito de reconocimiento y/o comprensión a lo largo del tiempo para ajustar la retroalimentación. Por ejemplo, si la primera retroalimentación proporcionada es “programadores de Java”, y la segunda es “James Gosling” (cocreador del lenguaje de programación Java), un tercer ejemplo de retroalimentación confirma “sí, Python lo consiguió” al reconocer el presente sistema una parte adicional de la entrada que indica otro lenguaje de programación.

En otras realizaciones, el grado de importancia es una puntuación de confianza recibida del convertidor de voz a texto, y la retroalimentación repite una parte de la entrada que el presente sistema determina que está por debajo de la puntuación de confianza. En contraste con el ejemplo expuesto anteriormente, este ejemplo describe un escenario donde el presente sistema proporciona retroalimentación repitiendo la parte de la entrada, porque el sistema ha determinado que el proceso de conversión de voz a texto puede haber arrojado resultados incorrectos. Otras realizaciones del grado de importancia pueden indicar que hay muchas variaciones fonéticas cercanas que coinciden con la entrada del usuario.

En el intercambio 401 de ejemplo, el usuario escribe y/o pronuncia una oración completa que contiene dos partes de la entrada que tienen altos grados de importancia: programadores de Java 409 y área Bolton 410. En algunas realizaciones, el presente sistema proporciona retroalimentación ofreciendo una repetición auditiva de la última frase “área de Bolton”, y también resalta visualmente las partes de interés (intercambio 402) para permitir una edición fácil por parte del usuario. El presente sistema recibe una segunda entrada del usuario (intercambio 403). Por ejemplo, el usuario hace clic o toca la segunda frase “área de Bolton” 410 para corregir un error de reconocimiento (“Boston” se ha reconocido erróneamente como “Bolton”). El presente sistema deduce que el usuario tiene la intención de modificar la primera entrada (intercambio 401) utilizando una segunda entrada. Por ejemplo, el presente sistema deduce que el usuario tiene la intención de modificar la primera entrada determinando una similitud entre la segunda entrada y la parte de la entrada correspondiente con la primera. La similitud puede basarse en la detección de caracteres similares de la segunda entrada y la parte de la entrada correspondiente, o en la detección de variaciones similares, tales como las variaciones fonéticas entre la segunda entrada y la parte de la entrada correspondiente. En algunas realizaciones, el usuario corrige el error de reconocimiento pronunciando la segunda entrada 411 “área de Boston” nuevamente, o eligiendo de una lista de variantes de “Bolton”, la variante “Boston”. El presente sistema puede determinar una consulta alternativa utilizando la segunda entrada. Incluso en el caso de que la segunda entrada implique una interacción visual, el usuario no tiene que usar procedimientos tradicionales de corrección de texto y no tiene que esforzarse para colocar el punto de inserción al final de la palabra “Bolton” para corregir el error de reconocimiento. En cambio, el presente sistema permite al usuario tocar el área resaltada alrededor de la parte identificada de la entrada “área de Bolton”. El presente sistema permite al usuario editar la parte de la entrada usando una segunda entrada, ya sea pronunciando de nuevo la parte de la entrada o escribiendo la parte de la entrada “Boston”.

El presente sistema selecciona un subconjunto de elementos de contenido relevantes, con base en la comparación de la consulta alternativa y los metadatos asociados con el subconjunto de elementos de contenido. Por ejemplo, el presente sistema responde con un resumen de los resultados (intercambio 404) de la consulta alternativa. Los resultados se basan en metadatos asociados, como los que identifican a los programadores de Java. En algunas realizaciones, la respuesta incluye una pantalla visual combinada con una versión oral o auditiva de los elementos de contenido relevantes. El usuario puede proporcionar información posterior, como “que también tengan conocimientos de Python” (intercambio 405). En algunas realizaciones, el usuario comienza a hablar incluso antes de que el presente sistema haya terminado de proporcionar los resultados de que hay más de cien programadores de Java. Al escuchar que el usuario comienza a hablar (p. ej., en función de la actividad del micrófono), algunas realizaciones suspenden inmediatamente la reproducción de la respuesta. Esta situación es análoga a las conversaciones entre humanos, donde un hablante puede detenerse cuando alguien le interrumpe. El presente sistema deduce que la segunda entrada está destinada a complementar la entrada existente. El presente sistema responde con un subconjunto revisado de elementos de contenido, como un resumen de treinta y tres programadores para la nueva consulta (intercambio 406).

Como se ha descrito anteriormente, el presente sistema deduce que el usuario tenía la intención de complementar la consulta existente. El presente sistema añade la consulta recién pronunciada al texto existente que se muestra en la pantalla, con la frase recién añadida Python resaltada (intercambio 412). En algunas realizaciones, el presente sistema permite al usuario usar la interfaz de la barra de búsqueda para continuar la conversación. El usuario puede utilizar

cómodamente una interfaz de búsqueda existente para la búsqueda de palabras clave, no solo para escribir la entrada en lenguaje natural, sino también para mantener una conversación como si el usuario simplemente estuviera añadiendo una palabra adicional en la búsqueda incremental basada en texto. Como se ha descrito anteriormente, algunas realizaciones rastrean disfluencias tales como las pausas en la recepción de la entrada de texto y desglosan una secuencia de entrada (ya sea escrita como entrada de texto, una entrada de voz convertida a texto o una combinación), en oraciones y frases basadas en la información de las pausas y las interacciones anteriores del usuario con la interfaz.

Además, la presente interfaz de usuario permite que los usuarios que ya se sienten cómodos con la búsqueda de palabras clave utilicen la interfaz familiar, sin tener que ser conscientes de que la interfaz es capaz de introducir información hablada y en lenguaje natural. Por ejemplo, después de escuchar el resumen de la sinopsis del subconjunto de elementos de contenido (intercambio 406), el usuario pronuncia una acción como “mándaselo por correo electrónico a Sam” (intercambio 407). El presente sistema identifica a “Sam” como una parte de la entrada basada en un grado de importancia. El presente sistema proporciona retroalimentación al usuario que identifica la parte de la entrada. Por ejemplo, el presente sistema repite “Sam” al usuario. El presente sistema también resalta la parte de la entrada “Sam”, lo que permite al usuario editar la parte de la entrada resaltada si es incorrecta. El presente sistema deduce que el usuario tenía la intención de complementar el intercambio conversacional anterior y envía la lista de treinta y tres programadores por correo electrónico a Sam. El sistema envía la lista por correo electrónico después de una pausa, para permitir al usuario editar la parte de entrada “Sam” si es necesario. Este ejemplo ilustra una secuencia de interacción que incluye entradas y salidas auditivas y visuales. Como se ha descrito anteriormente, el presente sistema no requiere tanto entradas y salidas auditivas como visuales; cualquier subconjunto o combinación de estas entradas y salidas es posible en función de la implementación en particular y el escenario de uso.

La figura 5 ilustra un ejemplo de interacción conversacional 500 entre el presente sistema y un usuario, según algunas realizaciones. La interacción conversacional 500 ilustra un ejemplo donde la interacción del usuario no implica la visualización para la entrada ni para la salida. En otras palabras, la interacción conversacional 500 ilustra el audio como el único medio de comunicación.

El usuario proporciona una primera entrada, “dime el nombre de Jessica Chastain” y hace una pausa para confirmar que el presente sistema ha identificado correctamente la parte de la entrada de interés para el usuario (intercambio 501). El presente sistema identifica que la parte de la entrada “Jessica Chastain” tiene un grado de importancia que excede un valor umbral. En consecuencia, el presente sistema proporciona retroalimentación al usuario que identifica la parte de la entrada. Por ejemplo, el presente sistema responde con “Sí, Jessica Chastain, continúa...” (intercambio 502). La retroalimentación proporcionada por el sistema representa una respuesta de audio natural, dado que la primera entrada del usuario es parcial y una oración incompleta (en contraste con el usuario que simplemente escribe una palabra clave) y dado que la primera entrada es una entrada de voz. (Algunas realizaciones rastrean si la fuente de la primera entrada es texto o voz, pese a la conversión de voz a texto de la entrada de voz). El usuario proporciona una entrada posterior “en su última película” (intercambio 503). El presente sistema deduce que el usuario tenía la intención de complementar la primera entrada con la segunda. El presente sistema determina una consulta alternativa combinando la primera entrada con la segunda, y selecciona un subconjunto de elementos de contenido comparando la consulta alternativa con los metadatos correspondientes. En función del subconjunto de elementos de contenido seleccionado, el presente sistema responde “*La noche más oscura*” (intercambio 504), la película más reciente donde Jessica Chastain había actuado en el momento de la consulta del usuario. Algunas realizaciones utilizan reglas gramaticales aplicadas a la estructura de la oración para determinar si la entrada es completa o es parcial. Por ejemplo, si el motor de reglas gramaticales determina que falta el sujeto, objeto directo y/o objeto indirecto de una oración, el presente sistema determina que la entrada es incremental o parcial. Otras realizaciones usan técnicas de clasificación probabilística, como un clasificador bayesiano ingenuo, para determinar cuándo una entrada está incompleta. Aun así, otras realizaciones deducen que un intercambio está incompleto en función de las relaciones entre entidades detectadas en el intercambio. Por ejemplo, el presente sistema deduce que el intercambio 501 está incompleto en función de las relaciones detectadas entre entidades gramaticales en el intercambio, porque no tiene sentido que un usuario pregunte el nombre (“dígame el nombre”) de alguien a quien acaba de nombrar (“de Jessica Chastain”).

El usuario hace una pregunta de seguimiento, “en qué película actuó que fue dirigida por Terrence”, y hace una pausa para recordar el nombre completo del director (intercambio 505). El presente sistema determina que una parte de la entrada (“Terrence”) tiene un grado de importancia que excede un umbral, en función de la disfluencia detectada (p. ej., en función de la pausa del usuario). El presente sistema determina además que el usuario tenía la intención de complementar la consulta existente, y determina una consulta alternativa que combina el término “Terrence” con la consulta existente (p. ej., película de Jessica Chastain). El presente sistema selecciona un subconjunto de elementos de contenido basándose en la comparación de la consulta alternativa y los metadatos correspondientes. Al descubrir una fuerte coincidencia inequívoca (p. ej., tanto para la parte de la entrada reconocida (“Terrence”) como para el subconjunto seleccionado de elementos de contenido (p. ej., película “*El árbol de la vida*”)) que califica un índice de éxito de respuesta por encima de un umbral, el presente sistema interrumpe al usuario para presentar el subconjunto

de elementos de contenido. Por ejemplo, el presente sistema exclama “te refieres a *El árbol de la vida* de Terrence Malick” (intercambio 506). En algunas realizaciones, el índice de éxito de la respuesta para un subconjunto coincidente de elementos de contenido se determina en función de factores que incluyen recuentos de ambigüedades en la entrada del usuario, recuentos de correcciones de errores necesarios para llegar a una coincidencia, recuentos de “saltos” entre nodos en una estructura de datos gráficos que representa el contexto conversacional para llegar a la coincidencia (en realizaciones que usan información representada en formato teórico de gráficos, como se describe a continuación), y/o un grado de relación de la coincidencia con interacciones conversacionales anteriores entre el sistema y el usuario. En realizaciones adicionales, el intercambio 506, aunque se ilustra en forma de audio, también puede producirse como una combinación de texto y audio. Además, como se ha descrito anteriormente, la entrada del usuario puede ser una cadena de palabras clave proporcionadas por el usuario en forma de texto. El presente sistema responde con la misma interjección incluso en respuesta a la entrada de texto. En otras realizaciones, la interjección es pasiva, ya que el presente sistema muestra los resultados en cuanto el usuario escribe “Terrence”.

Mientras que en el escenario de ejemplo descrito anteriormente, el presente sistema presenta de manera proactiva el subconjunto de elementos de contenido al usuario al determinar que la intención del usuario no era ambigua, la presente interfaz de voz incremental también permite la desambiguación de la entrada incluso antes de que el usuario termine de expresar una intención por completo. Por ejemplo, si el usuario tiene la intención de hacer la pregunta “juegan los sox esta noche” y dice “juegan los sox” y hace una pausa para escuchar los comentarios sobre la parte de la entrada “sox”, el presente sistema proporciona retroalimentación desambiguando automáticamente las entradas recibidas como los equipos de béisbol “Boston Red Sox” y “Chicago White Sox”. Si el usuario hubiera pronunciado completamente la oración, “juegan los sox esta noche”, y suponiendo que el presente sistema no estuviera al tanto de las preferencias personales del usuario, proporcionaría retroalimentación utilizando una pregunta de seguimiento para desambiguar la entrada recibida. Por ejemplo, el presente sistema pregunta “¿te refieres a los Boston Red Sox o a los Chicago White Sox?” Como se ha descrito anteriormente, el presente sistema puede proporcionar retroalimentación tanto de forma visual como auditiva. Además, en la presente búsqueda incremental, la desambiguación puede ocurrir cuando el usuario está expresando su intención. Por ejemplo, el presente sistema utiliza una disfluencia (p. ej., una pausa) proporcionada por el usuario en los límites de la frase para proporcionar retroalimentación en forma de confirmación al usuario de que el sistema recibió la conversión correcta de voz a texto en las frases de entrada. El presente sistema utiliza además una disfluencia detectada para identificar una parte de la entrada para desambiguar cuando la entrada tiene ambigüedades. En otras palabras, el usuario usa pausas en los límites de la frase para confirmar que el presente sistema ha entendido la entrada correctamente (con la capacidad de editar la entrada inmediatamente si se ha entendido de manera incorrecta). El usuario además utiliza pausas en los límites de la frase para desambiguar la intención. Como se ha ilustrado anteriormente, cuando el presente sistema determina que un índice de éxito de respuesta esperado supera un umbral, las disfluencias detectadas eliminan incluso la necesidad de que el usuario exprese más su intención. Según lo descrito anteriormente, las técnicas expuestas en la patente de EE. UU. n.º 2014/0039895 A1 se pueden utilizar para deducir el significado de la pausa de un usuario en el discurso y/o cualquier otra disfluencia de la voz. Por ejemplo, si se produce una pausa después de una entrada de oración parcial, el presente sistema deduce que el usuario busca confirmación de que se le ha entendido. Por el contrario, si la pausa precede a lo que se puede predecir como una parte de la entrada, el presente sistema deduce que el usuario no está seguro durante la entrada, y el sistema pondera en consecuencia el correspondiente grado de importancia de la parte de la entrada.

Detección y uso de las disfluencias

Las figuras 6-8 ilustran ejemplos de muestras de formas de onda de la entrada del usuario “Quién actuaba en la película El mañana nunca muere”, según algunas realizaciones.

<duración de la pausa = 800 ms>quién actuaba en la película<duración de la pausa = 550 ms>el mañana nunca muere<duración de la pausa = 1200 ms>

La entrada de voz 601 está flanqueada por una pausa inicial/silencio 602 y una pausa final/silencio 603. Además, dentro de la entrada de voz 601, hay una pausa 701 de 550 milisegundos. Estas pausas y/o silencios están indicados por una baja intensidad de la onda de sonido de la entrada de voz. A diferencia de estas pausas, la parte del discurso 801 tiene una alta intensidad, lo que indica que la parte del discurso 801 no es una disfluencia ni una pausa. Una definición del término pausa, como se usa en esta invención, es un período de silencio relativo donde el usuario no está hablando, pero donde la entrada de audio puede incluir sonidos ambientales. Por ejemplo, el presente sistema puede analizar los espectros de potencia de frecuencia 604, 704 y 804 para detectar el discurso frente a una pausa según los niveles de potencia de entrada. Como se muestra en los espectros de potencia 704, 804, la pausa 701 tiene una intensidad de aproximadamente -60 dB, y la parte de discurso 801 tiene una intensidad de aproximadamente -50 dB. Como la unidad de decibelios (dB) es una unidad logarítmica, hay un factor de diferencia de 10 en la intensidad de la pausa y la parte del discurso. En algunas realizaciones, los motores de voz a texto estándar realizan la detección de las disfluencias, teniendo en cuenta los sonidos ambientales.

Como se ha ilustrado anteriormente, se detecta una pausa cuando hay un período de ausencia o baja intensidad de sonido. La intensidad de corte del sonido para distinguir una pausa de una parte vocalizada de la entrada de voz se puede predefinir, por ejemplo, a -55 dB. Por otro lado, la intensidad de corte puede ser relativa a la entrada de voz y al ruido de fondo. La intensidad de corte se puede elegir, por ejemplo, al 20 % de la intensidad media de la entrada de voz. Si el ruido de fondo es alto, la intensidad de corte se puede elegir al 30 % de la intensidad media. Además, se puede predefinir el período mínimo de pausa de baja intensidad de sonido que forma una pausa. Por ejemplo, el período de pausa mínimo puede ser de 300 ms. Alternativamente, el período de pausa mínimo puede variar según la velocidad a la que el usuario hable. Si la entrada de voz se pronuncia rápido, el período de pausa mínimo puede ser más corto. Si la entrada de voz se pronuncia despacio, el período de pausa mínimo puede ser más largo. Por lo tanto, el presente sistema detecta una pausa cuando hay un período más largo que el período de pausa mínimo con una intensidad de sonido menor que la intensidad de corte.

Los motores de voz a texto tradicionales pueden intentar determinar palabras y/o frases basadas en la entrada de audio durante la pausa, o simplemente pueden detener el procesamiento del lenguaje durante la pausa. Una distinción de las realizaciones descritas en la presente invención es que las técnicas actualmente descritas incluyen el hecho de que la pausa se produjo como entrada a los módulos posteriores al motor de voz a texto para determinar la intención del usuario o para ayudar al usuario a formular la solicitud de consulta en sí. Además del uso de pausas, se utilizan otras formas de disfluencias, incluidos los rellenos de tiempo auditivos, en el procesamiento del habla. En caso de que el usuario pronuncie palabras o sonidos de relleno añadidos para acompañar una pausa, los módulos posteriores que procesan la salida del motor de voz a texto pueden reconocer esas palabras y sonidos de relleno añadidos a la pausa. Por ejemplo, el uso de palabras de relleno del tipo “como” seguidas de una pausa, o sonidos como “umm”, “hmm”, “bueno”, “uh” y “eh” seguidos de una pausa también se consideran en su conjunto como una pausa con la duración total de la pausa, incluida la duración de la pronunciación de las palabras de relleno. En otras realizaciones, las palabras de relleno auditivo no van seguidas de una pausa. Por lo general, los rellenos de tiempo auditivo son continuos y carecen de variaciones en el tono y el volumen. Estas características pueden ayudar a la detección de los rellenos de tiempo auditivos.

Si la entrada de voz del usuario, por otro lado, fue “¿Juegan los Red Sox mañana?”, es poco probable que haya una latencia cognitiva de recuperación que preceda a la palabra “mañana”, puesto que la instancia de la parte de la entrada “mañana” forma parte de la razón misma para hacer la pregunta. En contraste, durante la entrada de voz “¿Quién actuaba en (pausa) El mañana nunca muere?” el usuario puede hacer una pausa antes de “mañana” para demarcar conscientemente el límite de la frase (*es decir,* para identificar la parte de la frase “El mañana nunca muere” como un elemento distinto) o simplemente hacer una pausa para realizar un recuerdo cognitivo. Al utilizar la pausa que precede a la frase “El mañana nunca muere” para identificar la frase como una parte de la entrada, el presente sistema de recuperación de información puede comprender mejor que la intención del usuario se refiere a esa parte de la entrada. Esta valiosa información se puede utilizar para restringir la búsqueda a información que se refiere solo a esa parte de la entrada, o que las partes de la entrada devueltas por la búsqueda que están relacionadas con la película “El mañana nunca muere” pueden recibir un mayor peso de relevancia.

En el caso de demarcar el límite de la frase, el usuario puede decir con confianza la parte que sigue a la pausa. En consecuencia, el presente sistema puede determinar la parte que sigue a la pausa como una determinada frase o título en función del volumen o la velocidad de la voz del hablante. Otro procedimiento para distinguir si la parte que sigue a la pausa es una frase pronunciada con o sin confianza podría ser basarse donde se exprese adicionalmente después de la pausa inicial. Si una persona no está segura de una frase, es posible que haga una nueva pausa. Además, una pausa seguida de una frase dicha con confianza puede ser relativamente corta. Por lo tanto, el sistema puede suponer primero que una frase o título que siga a una pausa breve es una frase dicha con confianza. A continuación, el sistema realiza la búsqueda, pero si no encuentra ningún resultado, puede deducir que la frase que sigue a la pausa breve se ha dicho sin confianza.

Como se ha mencionado anteriormente, la presencia de una pausa dentro de la entrada de voz se puede usar como una forma de medir la confianza de las partes de la propia entrada. La interpretación de la duración de las pausas y la frecuencia con que ocurren también se tiene en cuenta en las realizaciones de la presente invención para distinguir los casos de usuarios que simplemente hablan despacio (para que el reconocimiento de voz funcione mejor) frente a las pausas para realizar el recuerdo cognitivo. Por ejemplo, supongamos que la entrada de voz del usuario fue “¿Quién actuaba en (pausa) El día nunca muere?” En este caso, el sistema puede usar la pausa para indicar que el usuario puede no estar seguro del nombre del elemento para el que solicita información. Por lo tanto, cuando no encuentra un elemento correspondiente a “El día nunca muere”, el sistema puede responder con preguntas, orientado por la entrada del usuario (utilizando, p. ej., las técnicas establecidas en las aplicaciones incorporadas como se ha descrito anteriormente) para ayudar al usuario a definir su intención.

Además, el presente sistema puede dar una prioridad de búsqueda menor al elemento expresado con poca confianza

en su conjunto y, en su lugar, usar los elementos de mayor confianza para guiar la búsqueda. Por ejemplo, el sistema puede confiar mucho en la parte “Quién actuaba” para centrarse en los resultados de un dominio de entretenimiento de audio/vídeo (basándose en la palabra “actuaba”). Conociendo este dominio, el sistema puede refinar aún más la búsqueda basándose en las partes de la entrada de menor confianza. Por ejemplo, el sistema puede realizar consultas basadas en combinaciones de las palabras de la parte de menor confianza para encontrar lo que el usuario está buscando o al menos para proporcionarle algunas opciones al usuario. De este modo, el sistema puede responder con la afirmación de que no puede encontrar una película titulada “El día nunca muere” y preguntar si el usuario quiso decir “El amor nunca muere” o “El mañana nunca muere”.

Las figuras 9-10 ilustran formas de onda de ejemplo que muestran el uso de la falta de pausas para detectar el recorte en la entrada de voz, según algunas realizaciones. Específicamente, la figura 9 ilustra el recorte inicial 901 y la figura 10 ilustra el recorte final 1002. El recorte de voz inicial 901 y el recorte final 1002 son detectados por el motor de voz a texto en combinación con los otros módulos y se codifican junto con la entrada de voz como se ha ilustrado anteriormente. Por el contrario, la presencia de la pausa final 902 y la pausa inicial 1001 delimitan claramente la entrada de voz. Los motores de voz a texto tradicionales pueden asignar sonidos recortados a palabras que coinciden aproximadamente o simplemente emitir un texto fonético equivalente a los sonidos. Las implementaciones de la presente invención reconocen la ausencia de estas pausas delimitadoras y utilizan su presencia como información adicional para interpretar el significado de la entrada del usuario. Por ejemplo, en lugar de simplemente encontrar la palabra que más se acerca a la parte recortada 901, la realización ilustrativa considera la posibilidad de que el usuario tuviera la intención de decir una palabra diferente que tiene un sufijo que coincide.

Repositorios de información

En algunas realizaciones, la presente invención utiliza repositorios de información para buscar el resultado de la consulta o para encontrar una palabra o frase sustituta. Los repositorios de información están asociados con dominios, que son conjuntos de tipos similares de información y/o determinados tipos de elementos de contenido. Ciertos tipos de repositorios de información incluyen entidades y relaciones entre las entidades. Cada entidad/relación pertenece a un tipo, respectivamente, de un conjunto de tipos. Además, a cada entidad/relación se le asocia un conjunto de atributos, que pueden capturarse, en algunas realizaciones, como un conjunto finito definido de campos de nombre-valor. La asignación entidad/relación también sirve como un conjunto de metadatos asociados con los elementos de contenido porque la asignación entidad/relación proporciona información que describe los diversos elementos de contenido. En otras palabras, una entidad en particular tendrá relación con otras entidades y estas “otras entidades” servirán como metadatos para la “entidad en particular”. Además, en la asignación, cada entidad puede tener atributos asignados a ella o a las relaciones que conectan la entidad con otras entidades. Colectivamente, esto constituye los metadatos asociados con las entidades/elementos de contenido. En general, dichos repositorios de información pueden denominarse en la presente invención *repositorios de información estructurada.* A continuación se presentan ejemplos de repositorios de información asociados a los dominios.

En un dominio **de entretenimiento de medios** se incluyen entidades, tales como películas, programas de televisión, episodios, equipo de producción, roles/personajes, actores/personalidades, atletas, partidos, equipos, ligas y torneos, deportistas, artistas e intérpretes de música, compositores, álbumes, canciones, personalidades de actualidad y/o distribuidores de contenido. Estas entidades tienen una serie de relaciones que se capturan en el repositorio de información. Por ejemplo, una entidad de una película se relaciona con una o más entidades de actor/personalidad mediante la relación “actuaba en”. De manera similar, una entidad de película puede estar relacionada con una entidad de álbum de música a través de la relación “banda sonora original”, que a su vez puede estar relacionada con una entidad de canción a través de la relación de “pista de un álbum”. Mientras tanto, los nombres, las descripciones, la información de la programación, las reseñas, las calificaciones, los costes, las URL de los vídeos o audios, las transacciones de la tienda de contenidos o aplicaciones, las puntuaciones, etc., pueden considerarse campos de atributo.

En un **dominio de correo electrónico (e-mail) personal** se incluyen entidades, como correos electrónicos, hilos de correo electrónico, contactos, remitentes, destinatarios, nombres de compañías, departamentos/unidades de negocios de empresa, carpetas de correo electrónico, ubicaciones de oficinas y/o ciudades y países correspondientes a las ubicaciones de las oficinas. Ejemplos ilustrativos de las relaciones serían una entidad de correo electrónico relacionada con la entidad remitente (así como las entidades para, cc, cco, receptores e hilo de correo electrónico). Mientras tanto, pueden existir relaciones entre un contacto y su empresa, departamento, ubicación de la oficina. En este repositorio, ejemplos de los campos de atributo asociados con las entidades serían los nombres de contactos, designaciones, identificadores de correo electrónico, otra información de contacto, marca de tiempo de correo electrónico enviado/recibido, asunto, cuerpo, archivos adjuntos, niveles de prioridad, información de ubicación de una oficina y/o el nombre y descripción de un departamento.

En un dominio **relacionado con viajes**/**hoteles y turismo** se incluyen entidades, como ciudades, hoteles, marcas de

hoteles, puntos de interés individuales, categorías de puntos de interés, cadenas minoristas de cara al consumidor, sitios de alquiler de automóviles y/o empresas de alquiler de automóviles. Entre las relaciones entre tales entidades se incluyen la ubicación, pertenencia a cadenas y/o categorías. También se incluyen en los campos de atributo nombres, descripciones, palabras clave, costes, tipos de servicio, calificaciones, reseñas, etc.

En un **dominio de comercio electrónico** se incluyen entidades como artículos de productos, categorías y subcategorías de productos, marcas, tiendas, etc. Las relaciones entre dichas entidades pueden incluir información de compatibilidad entre artículos de productos, un producto “vendido por” una tienda, etc. En los campos de atributo se incluyen las descripciones, palabras clave, reseñas, calificaciones, costes y/o información de disponibilidad.

En un **dominio de libro de direcciones** se incluyen entidades e información como nombres de contacto, direcciones de correo electrónico, números de teléfono, direcciones físicas y empresa.

Las entidades, relaciones y atributos enumerados aquí son solo ilustrativos y no pretenden ser una lista exhaustiva.

Algunas realizaciones también pueden usar repositorios que *no* son repositorios de información estructurada como los descritos anteriormente. Por ejemplo, el repositorio de información correspondiente a documentos basados en la red (*p. ej*., Internet/World Wide Web) puede considerarse una red de relaciones web de documentos vinculados (entidades). Sin embargo, en general, ninguna estructura de tipos directamente aplicable puede describir de manera significativa, de forma no trivial, todos los tipos de entidades y relaciones y atributos asociados con los elementos de Internet en el sentido de los repositorios de información estructurada descritos anteriormente. Sin embargo, elementos como los nombres de dominio, tipos de medios de Internet, nombres de archivo, extensión de nombre de archivo, etc. pueden usarse como entidades o atributos con dicha información.

Por ejemplo, consideremos un corpus que conste de un conjunto de documentos de texto no estructurados. En este caso, ninguna estructura de tipos directamente aplicable puede enumerar un conjunto de entidades y relaciones que describan de manera significativa el contenido del documento. Sin embargo, la aplicación de técnicas de procesamiento de extracción semántica de la información como un paso de preprocesamiento puede dar como resultado entidades y relaciones que pueden descubrir parcialmente la estructura de dicho corpus.

Ejemplos ilustrativos de acceso a repositorios de información.

La siguiente descripción ilustra algunos ejemplos de tareas de recuperación de la información en el contexto de los repositorios de información estructurados y no estructurados como se ha descrito anteriormente.

En algunos casos, un usuario está interesado en una o más entidades de algún tipo, generalmente llamado tipo de intención en la presente invención, que el usuario desea descubrir especificando solo las restricciones de campo de atributo con las que las entidades deben cumplir. Tenga en cuenta que a veces la intención puede ser doble (tipo, atributo) cuando el usuario desea algún atributo de una entidad de cierto tipo. Por ejemplo, si el usuario desea conocer la calificación de una película, la intención podría verse como (tipo, atributo) = (película, calificación). Tales restricciones de la consulta se denominan generalmente *restricciones de solo atributo* en la presente invención.

Cada vez que el usuario nombra la entidad o especifica suficiente información para que coincida directamente con los atributos de la entidad de tipo de intención deseada, se trata de una restricción de solo atributo. Por ejemplo, cuando el usuario identifica una película por su nombre y algún atributo adicional (*p. ej*., *El cabo del miedo* de los años 60), o cuando especifica una coincidencia de asunto para el correo electrónico que desea descubrir, o cuando solicita hoteles en función de un rango de precios, o cuando especifica que quiere un iPod touch de color negro de 32 GB.

Sin embargo, en algunos casos, un usuario está interesado en una o más entidades del tipo de intención y especifica no solo las restricciones de campo de atributo a las entidades de tipo de intención, sino también restricciones de campo de atributo o incluso nombra *otras entidades* con las que las entidades de tipo de intención están *conectadas* a través de las relaciones de alguna manera bien definida. A tales restricciones de la consulta se las denomina generalmente *restricciones orientadas a la conexión* en la presente invención.

Un ejemplo de restricción orientada a la conexión sería cuando el usuario quiere conocer una película (un tipo de intención) basada en la especificación de dos o más actores de la película o una película sobre un actor y un premio que ganó la película. Otro ejemplo, en el contexto del correo electrónico, es si el usuario desea ver los correos electrónicos (tipo de intención) recibidos de ciertos remitentes de una empresa en particular en los últimos siete días. Del mismo modo, otro ejemplo es si el usuario desea reservar una habitación de hotel (tipo de intención) cercana a una estación de tren y también a un Starbucks. Otro ejemplo es si el usuario quiere encontrar un televisor (tipo de intención) fabricado por Samsung que también sea compatible con una NINTENDO WII. Todos estos son ejemplos de consultas de restricciones orientadas a la conexión.

En los anteriores ejemplos de restricciones orientadas a la conexión, el usuario describe o especifica explícitamente las otras entidades conectadas a las entidades de intención. A tales restricciones se las denomina generalmente en la presente invención *restricciones explícitas orientadas a la conexión* y a las entidades mencionadas en la presente invención se las denomina *entidades explícitas.*

Mientras tanto, otras consultas contienen restricciones orientadas a la conexión que incluyen entidades no especificadas o implícitas como parte de la especificación de la restricción. En tal situación, el usuario está intentando identificar una información, entidad, atributo, etc. que no se conoce a través de las relaciones entre el elemento desconocido y los elementos que el usuario sí conoce. A tales restricciones se las denomina generalmente en la presente invención *restricciones implícitas orientadas a la conexión* y a las entidades no especificadas se las denomina generalmente en la presente invención *entidades implícitas* de la restricción.

Por ejemplo, el usuario puede desear identificar una película que está buscando nombrando dos personajes de la misma. Sin embargo, el usuario no recuerda el nombre de uno de los personajes, pero sí recuerda que un actor en particular interpretó al personaje. Por lo tanto, en su consulta, hace referencia a un personaje por su nombre e identifica al personaje desconocido indicando que el personaje fue interpretado por el actor en particular.

Sin embargo, hay que tener en cuenta las siguientes restricciones de usuario para los objetivos específicos de recuperación de la información: El usuario quiere conocer el papel (intención) interpretado por una actriz específica (p. ej., “Michelle Pfeiffer”) en una película no especificada que trata sobre un papel en concreto (p. ej., el personaje “Tony Montana”). En este caso, la restricción del usuario incluye una entidad implícita o no especificada que corresponde a la película “Scarface”. Del mismo modo, supongamos que el usuario quiere conocer la película (intención) protagonizada por la actriz especificada “Scarlett Johannson” y el actor no especificado que interpretó el papel concreto de “Obe Wan Kanobi” en una película específica *Star Wars*. En este caso, la entidad implícita es el actor “Ewan McGregor” y la entidad intencional es la película *La isla* protagonizada por “Scarlett Johannson” e “Ewan McGregor”.

En el contexto del repositorio de correo electrónico, un ejemplo sería un usuario que desea obtener el último correo electrónico (intención) de un hombre no especificado de una compañía concreta “Intel” a quien se le presentó por correo electrónico (un especificador de atributos) la semana pasada. En este caso, la entidad implícita es un contacto que se puede descubrir examinando los contactos de “Intel”, a través de una relación empleado/empresa, que fue la primera vez que intercambió un correo electrónico normal con el usuario la semana pasada.

Los tres ejemplos anteriores son restricciones orientadas a la conexión, pero incluyen entidades no especificadas o implícitas como parte de la especificación de la restricción. La presente invención se refiere a las restricciones aquí contenidas como *restricciones implícitas orientadas a la conexión* y se refiere a las entidades no especificadas aquí contenidas como *entidades implícitas* de la restricción.

En el contexto de las restricciones orientadas a la conexión, puede ser útil asignar entidades y relaciones de los repositorios de información a los nodos y bordes de una estructura de datos teóricos de gráficos. La motivación para emplear un modelo gráfico en lugar de un modelo de relaciones de entidades es observar que la relevancia, la proximidad y la relación de la conversación en lenguaje natural pueden modelarse simplemente con nociones como la distancia entre enlaces y, en algunos casos, los caminos más cortos y los árboles de menor peso. Durante la conversación, cuando el diálogo de usuario involucra a otras entidades relacionadas con las entidades realmente buscadas, una recuperación de información de direccionamiento de subrutina como un problema simple de búsqueda de gráficos ayuda de manera eficaz a reducir la dependencia de una comprensión profunda e inequívoca de la estructura de la oración, lo que puede ser una enorme ventaja de la implementación. Incluso si el cálculo de la intención del usuario es ambiguo o no concluyente, siempre y cuando se hayan reconocido las entidades en el enunciado del usuario, un tratamiento del problema basado en la interpretación del gráfico permite que nuestro sistema responda de una manera mucho más inteligible que cualquier otra.

Algunas realizaciones de la presente invención utilizan una firma de preferencias del usuario (que captura la actividad e intereses del usuario, tanto implícita como explícitamente determinados) de una manera dependiente del contexto y, si corresponde, aplica la personalización a la selección de una palabra precedida por una pausa o un palabra recortada en caso de recorte inicial y recorte final. La personalización también se puede aplicar a la selección de resultados para ofrecer la mejor respuesta que tenga una alta probabilidad de coincidir con la intención del usuario. Ciertas realizaciones de la presente invención utilizan la firma de las preferencias del usuario, si están disponibles, para resolver la ambigüedad de la entrada del usuario.

El sistema puede proporcionar una firma de preferencia del usuario que utiliza técnicas para descubrir y almacenar dicha información de preferencias del usuario. Por ejemplo, los procedimientos y sistemas establecidos en la patente de EE. UU. n.º 7.774.294 denominada *Procedimientos y sistemas para seleccionar y presentar contenido basado en*

la periodicidad aprendida de las selecciones de contenido del usuario, concedida el 10 de agosto de 2010, la patente de EE. UU. n.º 7.835.998 denominada *Procedimientos y sistemas para seleccionar y presentar contenido en un primer sistema basado en las preferencias del usuario aprendidas en un segundo sistema,* concedida el 16 de noviembre de 2010, la patente de EE. UU. n.º 7.461.061 denominada *Procedimientos y sistemas de interfaz de usuario para seleccionar y presentar contenido basado en la navegación del usuario y las acciones de selección asociadas con el contenido,* concedida el 2 de diciembre de 2008, y la patente de EE. UU. n.º 8.112.454 denominada *Procedimientos y sistemas para ordenar elementos de contenido según las preferencias de usuario aprendidas,* concedida el 7 de febrero de 2012, se pueden usar con las técnicas aquí descritas. Sin embargo, el uso de firmas y/o información de preferencias del usuario no se limita a las técnicas establecidas en las aplicaciones incorporadas.

En la descripción anterior, ciertos pasos o procesos se pueden realizar en servidores particulares o como parte de un motor particular. Estas descripciones son meramente ilustrativas, ya que los pasos específicos se pueden realizar en varios dispositivos de hardware, incluidos, entre otros, sistemas de servidor y/o dispositivos móviles. De manera similar, la división de dónde se realizan los pasos en particular puede variar, entendiéndose que dentro del alcance de la invención no se encuentra ninguna división o una división diferente. Además, el uso de “analizador”, “módulo”, “motor” y/u otros términos utilizados para describir el procesamiento del sistema informático está destinado a ser intercambiable y representar la lógica o los circuitos donde se puede ejecutar la funcionalidad.

Las técnicas y sistemas descritos en esta invención pueden implementarse como un programa informático para su uso con un sistema informático o dispositivo electrónico computarizado. Dichas implementaciones pueden incluir una serie de instrucciones computarizadas, o lógica, fijadas en un medio tangible, como un medio legible por ordenador (por ejemplo, un disquete, CD-ROM, ROM, memoria flash u otra memoria o disco duro) o transmitible a un sistema informático o un dispositivo, a través de un módem u otro dispositivo de interfaz, como un adaptador de comunicaciones conectado a una red a través de un medio.

El medio puede ser un medio tangible (p. ej., las líneas de comunicaciones ópticas o analógicas) o un medio implementado con técnicas inalámbricas (p. ej., Wi-Fi, tecnología móvil, de microondas, infrarrojos u otras técnicas de transmisión). La serie de instrucciones computarizadas incorpora al menos parte de la funcionalidad descrita en este documento con respecto al sistema. Los expertos en la materia deberían apreciar que tales instrucciones computarizadas se pueden escribir en varios lenguajes de programación para su uso con muchas arquitecturas informáticas o sistemas operativos.

Además, dichas instrucciones pueden almacenarse en cualquier dispositivo de memoria tangible, como dispositivos de memoria semiconductores, magnéticos, ópticos y otros, y pueden transmitirse utilizando cualquier tecnología de comunicaciones, como óptica, infrarroja, de microondas u otras tecnologías de transmisión.

Se espera que dicho programa informático se distribuya como un medio extraíble con la documentación impresa o electrónica adjunta (p. ej., software precintado), precargado en un sistema informático (p. ej., en la ROM del sistema o en el disco duro), o distribuido desde un servidor o boletín de anuncios electrónico a través de la red (p. ej., Internet o World Wide Web). Por supuesto, algunas realizaciones de la invención pueden implementarse como una combinación de software (p. ej., un programa informático) y hardware. Aun así, otras realizaciones de la invención se implementan completamente como hardware o completamente como software (p. ej., un programa informático).

REIVINDICACIONES

1. Un procedimiento implementado por ordenador para seleccionar y presentar elementos de contenido basado en las entradas de los usuarios que comprende:

proporcionar acceso a un conjunto de elementos de contenido, estando dichos elementos de contenido asociados a metadatos que describen uno de los elementos de contenido;

recibir una primera entrada (401) destinada por el usuario a identificar al menos un elemento de contenido deseado;

determinar que al menos una parte de la primera entrada tiene un grado de importancia que excede un valor umbral;

proporcionar retroalimentación al usuario (402) identificando parte de la primera entrada;

recibir una segunda entrada (403, 405) del usuario posterior a la primera entrada;

deducir si el usuario tenía la intención de modificar la primera entrada con la segunda entrada o complementar la primera entrada con la segunda entrada;

con la condición de la que se deduce que el usuario tenía la intención de modificar la primera entrada con la segunda entrada, determinar una consulta alternativa modificando la primera entrada basada en la segunda entrada (404);

con la condición de la que se deduce que el usuario tenía la intención de complementar la primera entrada con la segunda entrada, determinar una consulta alternativa combinando la primera entrada basada en la segunda entrada (406);

seleccionar un subconjunto de elementos de contenido del conjunto de elementos de contenido basándose en la comparación de la consulta alternativa y los metadatos asociados con los elementos de contenido del subconjunto de elementos de contenido; y

presentar el subconjunto de elementos de contenido al usuario;

caracterizado porque

la deducción de si el usuario tenía la intención de modificar la primera entrada con la segunda entrada o complementar la primera entrada con la segunda entrada incluye:

determinar un grado de similitud entre la primera entrada y la segunda entrada;

con la condición de que el grado de similitud esté por encima de un umbral, deducir que el usuario tenía la intención de modificar la primera entrada; y

con la condición de que el grado de similitud esté por debajo de un umbral, deducir que el usuario tenía la intención de complementar la primera entrada.

2. El procedimiento según la reivindicación 1,
donde la determinación de que la parte de la primera entrada tiene el grado de importancia que excede el valor umbral incluye la identificación de uno o más límites de la frase en la entrada incremental, y

donde la identificación de uno o más límites de la frase se basa al menos en parte en al menos uno de los siguientes (a) una disfluencia identificada del usuario en la primera entrada, (b) reglas gramaticales aplicadas a la primera entrada, (c) el grado de importancia de la parte de la primera entrada, (d) al menos una interacción conversacional previa con el usuario, y (e) una firma de preferencia del usuario

la firma de preferencia del usuario describe las preferencias del usuario para al menos uno de (i) los elementos de contenido en particular y (ii) los metadatos en particular asociados con los elementos de contenido, donde la parte de la primera entrada se identifica en función de la firma de preferencia del usuario.

3. El procedimiento según la reivindicación 2, donde la disfluencia incluye al menos una pausa en la entrada de voz, un relleno de tiempo auditivo en la entrada de voz y una pausa en la entrada de escritura.

4. El procedimiento según la reivindicación 1, donde la selección del subconjunto de elementos de

contenido se basa además en una disfluencia identificada en la primera entrada, y también en interacciones conversacionales previas que se determina que están relacionadas con la primera entrada y la segunda entrada.

5. El procedimiento según la reivindicación 1, donde la retroalimentación proporcionada incluye al menos uno de los siguientes:

la solicitud de aclaración sobre la parte identificada de la primera entrada basada, al menos en parte, en una determinación de que se produce una primera disfluencia después de que el usuario haya proporcionado parte de la primera entrada,

la sugerencia de completar la primera entrada recibida basada, al menos en parte, en una determinación de que la segunda disfluencia ocurre antes de que se espere que el usuario proporcione la parte de la primera entrada, y

la repetición de la parte de la primera entrada al usuario, para notificarle que la parte de la primera entrada puede haberse reconocido incorrectamente.

6. El procedimiento según la reivindicación 1, donde la retroalimentación proporcionada al usuario se elige sobre la base de al menos uno de los siguientes:

la duración de la disfluencia identificada en la primera entrada,

un grado de confianza en que el reconocimiento de voz a texto de la parte de la primera entrada es correcto,

un recuento de las ambigüedades detectadas en la primera entrada,

un recuento de las correcciones de errores necesarias para identificar la parte de la primera entrada,

un recuento de nodos en una estructura de datos gráficos, donde la cantidad de nodos en la estructura de datos gráficos mide una ruta entre un primer nodo que representa un elemento de interés de una interacción conversacional previa y un segundo nodo que representa la parte de la primera entrada, y

un grado de relación entre la parte de la primera entrada y las interacciones conversacionales previas con el usuario.

7. Un sistema para seleccionar y presentar elementos de contenido basado en las entradas de los usuarios que comprende:

instrucciones legibles por ordenador codificadas en un medio legible por ordenador no transitorio, haciendo las instrucciones legibles por ordenador que el sistema informático esté configurado para:

proporcionar acceso a un conjunto de elementos de contenido, estando dichos elementos de contenido asociados a metadatos que describen uno de los elementos de contenido;

recibir una primera entrada (401) destinada por el usuario a identificar al menos un elemento de contenido deseado;

determinar que al menos una parte de la primera entrada tiene un grado de importancia que excede un valor umbral;

proporcionar retroalimentación al usuario (402) identificando parte de la primera entrada;

recibir una segunda entrada (403, 405) del usuario posterior a la primera entrada;

deducir si el usuario tenía la intención de modificar la primera entrada con la segunda entrada o complementar la primera entrada con la segunda entrada;

con la condición de la que se deduce que el usuario tenía la intención de modificar la primera entrada con la segunda entrada, determinar una consulta alternativa modificando la primera entrada basada en la segunda entrada (404);

con la condición de la que se deduce que el usuario tenía la intención de complementar la primera entrada con la segunda entrada, determinar una consulta alternativa combinando la primera entrada basada en la segunda entrada (406);

seleccionar un subconjunto de elementos de contenido del conjunto de elementos de contenido basándose en la comparación de la consulta alternativa y los metadatos asociados con los elementos de contenido del subconjunto de

elementos de contenido; y

presentar el subconjunto de elementos de contenido al usuario;

caracterizado porque

las instrucciones legibles por computadora hacen que el sistema se configure de modo que la deducción de si el usuario tenía la intención de modificar la primera entrada con la segunda entrada o complementar la primera entrada con la segunda entrada comprende:

determinar un grado de similitud entre la primera entrada y la segunda entrada;

con la condición de que el grado de similitud esté por encima de un umbral, deducir que el usuario tenía la intención de modificar la primera entrada; y

con la condición de que el grado de similitud esté por debajo de un umbral, deducir que el usuario tenía la intención de complementar la primera entrada.

8. El sistema según la reivindicación 7,
donde la determinación de que la parte de la primera entrada tiene el grado de importancia que excede el valor umbral incluye las instrucciones legibles por ordenador que hacen que el sistema informático esté configurado para identificar uno o más límites de la frase en la entrada incremental, y

donde la identificación de uno o más límites de la frase se basa al menos en parte en al menos uno de los siguientes (a) una disfluencia identificada del usuario en la primera entrada, donde la disfluencia incluye al menos una pausa en la entrada de voz, un relleno de tiempo auditivo en la entrada de voz y una pausa en la entrada de escritura, (b) reglas gramaticales aplicadas a la primera entrada, (c) el grado de importancia de la parte de la primera entrada, (d) al menos una interacción conversacional previa con el usuario, y (e) una firma de preferencia del usuario, y

donde la firma de preferencia del usuario describe las preferencias del usuario para al menos uno de (i) los elementos de contenido en particular y (ii) los metadatos en particular asociados con los elementos de contenido, donde la parte de la primera entrada se identifica en función de la firma de preferencia del usuario.

9. El sistema según la reivindicación 7, donde la selección del subconjunto de elementos de contenido se basa además en una disfluencia identificada en la primera entrada, y también en interacciones conversacionales previas que se determina que están relacionadas con la primera entrada y la segunda entrada.

10. El sistema según la reivindicación 7, donde las instrucciones legibles por ordenador que hacen que el sistema esté configurado para proporcionar la retroalimentación incluye al menos uno de los siguientes:

las instrucciones legibles por ordenador que hacen que el sistema esté configurado para solicitar una aclaración sobre la parte identificada de la primera entrada basada, al menos en parte, en una determinación de que se produce una primera disfluencia después de que el usuario haya proporcionado parte de la primera entrada,

las instrucciones legibles por ordenador que hacen que el sistema esté configurado para sugerir que se complete la primera entrada recibida basada, al menos en parte, en una determinación de que la segunda disfluencia ocurre antes de que se espere que el usuario proporcione la parte de la primera entrada, y

las instrucciones legibles por ordenador que hacen que el sistema esté configurado para repetir la parte de la primera entrada al usuario, para notificarle que la parte de la primera entrada puede haberse reconocido incorrectamente.

11. El sistema según la reivindicación 7, donde la retroalimentación proporcionada al usuario se elige sobre la base de al menos uno de los siguientes:

la duración de la disfluencia identificada en la primera entrada,

un grado de confianza en que el reconocimiento de voz a texto de la parte de la primera entrada es correcto,

un recuento de las ambigüedades detectadas en la primera entrada,

un recuento de las correcciones de errores necesarias para identificar la parte de la primera entrada,

un recuento de nodos en una estructura de datos gráficos, donde la cantidad de nodos en la estructura de datos gráficos mide una ruta entre un primer nodo que representa un elemento de interés de una interacción conversacional previa y un segundo nodo que representa la parte de la primera entrada, y

un grado de relación entre la parte de la primera entrada y las interacciones conversacionales previas con el usuario.

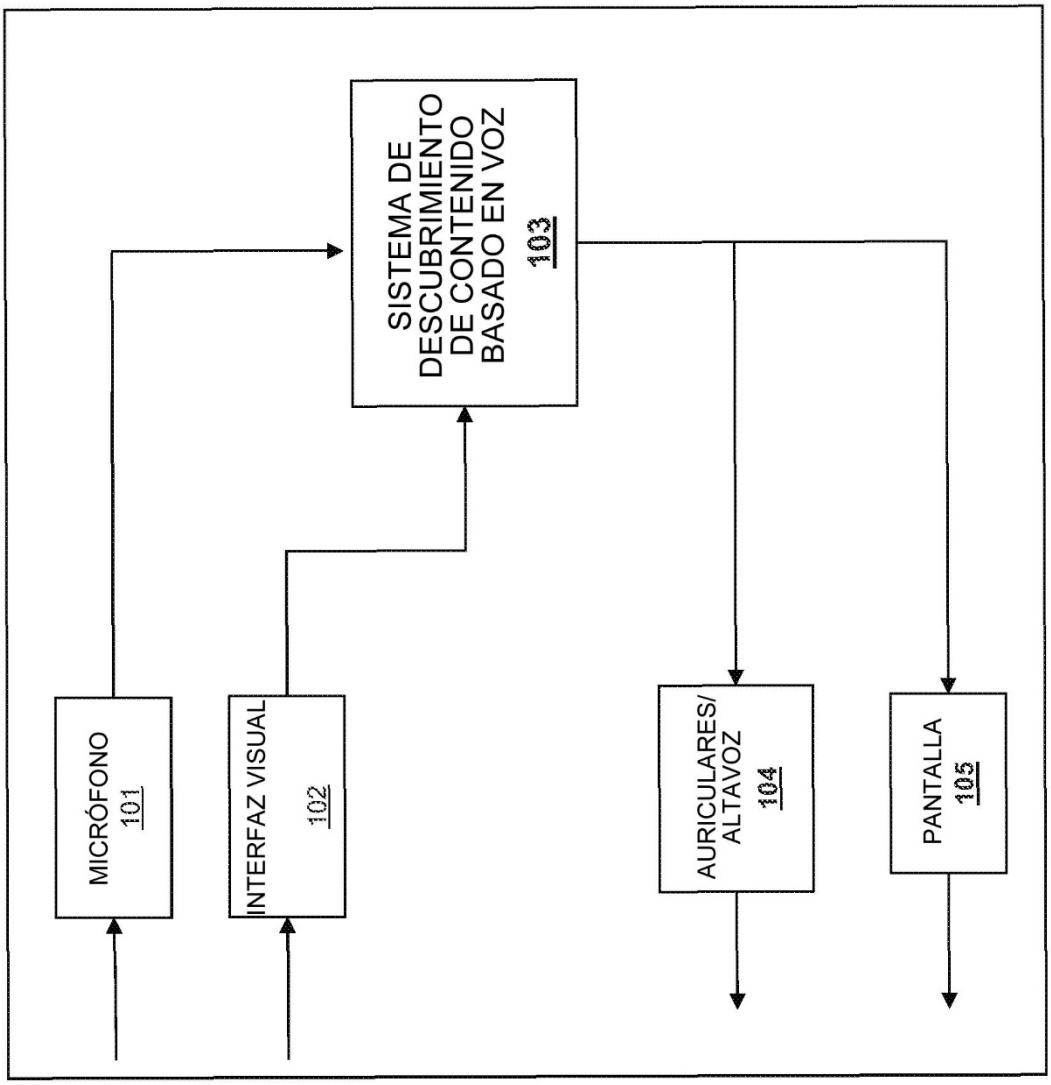
SISTEMA DE DESCUBRIMIENTO DE CONTENIDO BASADO EN VOZ
103
MICRÓFONO
101
INTERFAZ VISUAL
102
AURICULARES/ ALTAVOZ
104
PANTALLA
105

FIG.1

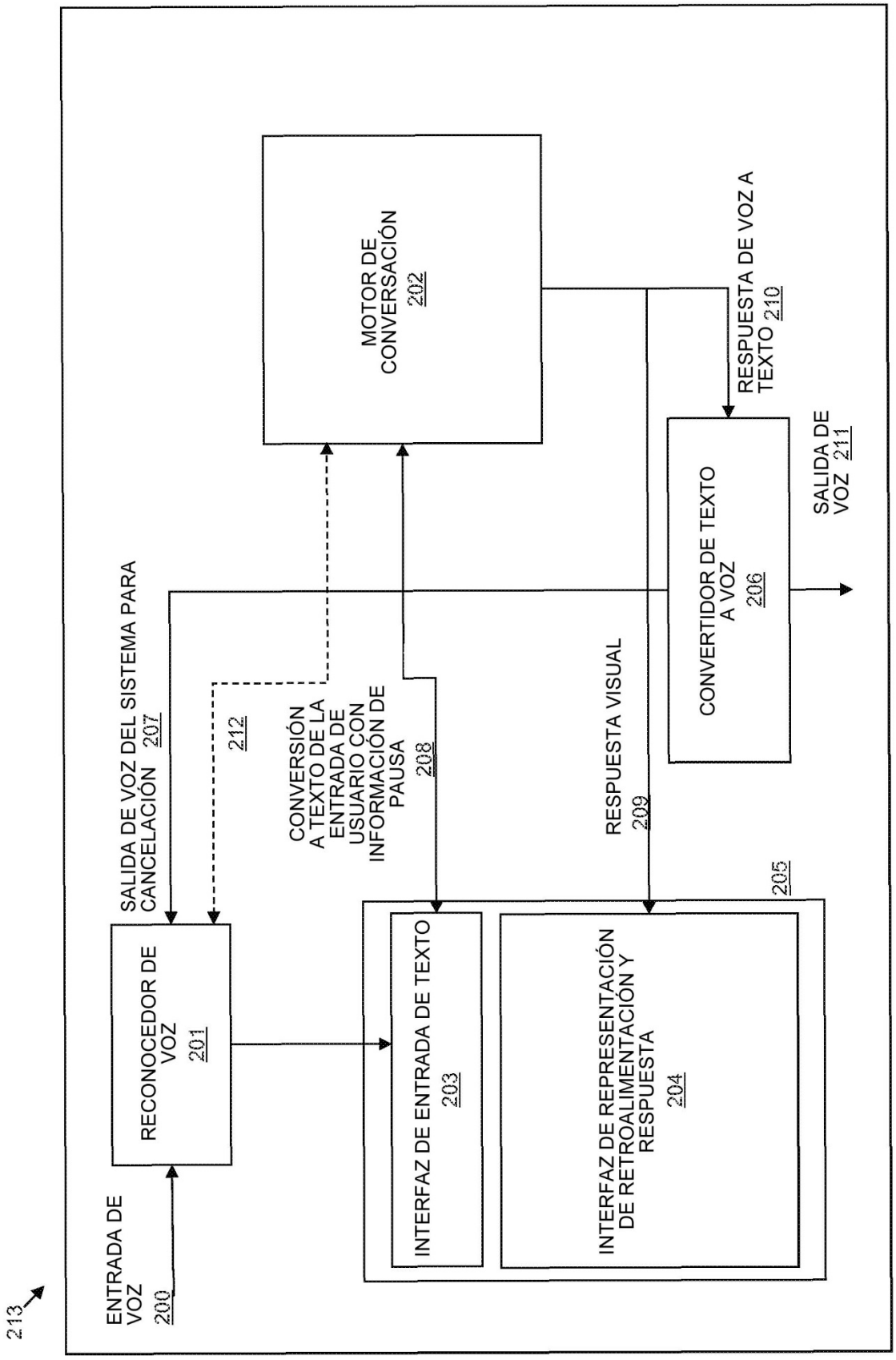
213
ENTRADA DE VOZ 200
RECONOCEDOR DE VOZ 201
SALIDA DE VOZ DEL SISTEMA PARA CANCELACIÓN 207
212
CONVERSIÓN A TEXTO DE LA ENTRADA DE USUARIO CON INFORMACIÓN DE PAUSA 208
MOTOR DE CONVERSACIÓN 202
205
INTERFAZ DE ENTRADA DE TEXTO 203
INTERFAZ DE REPRESENTACIÓN DE RETROALIMENTACIÓN Y RESPUESTA 204
RESPUESTA VISUAL 209
CONVERTIDOR DE TEXTO A VOZ 206
RESPUESTA DE VOZ A TEXTO 210
SALIDA DE VOZ 211

FIG. 2

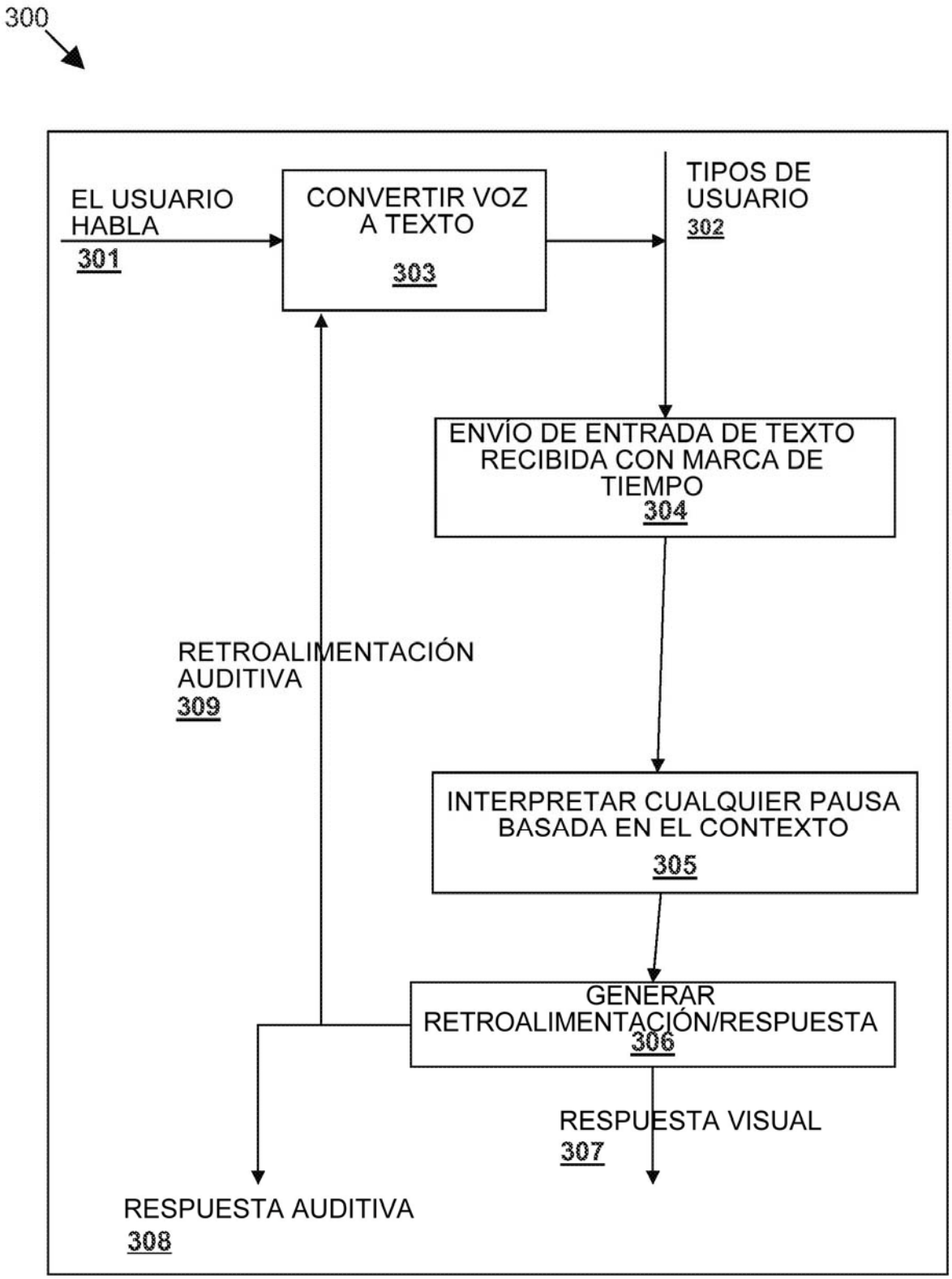
300
EL USUARIO HABLA
301
CONVERTIR VOZ A TEXTO
303
TIPOS DE USUARIO
302
ENVÍO DE ENTRADA DE TEXTO RECIBIDA CON MARCA DE TIEMPO
304
RETROALIMENTACIÓN AUDITIVA
309
INTERPRETAR CUALQUIER PAUSA BASADA EN EL CONTEXTO
305
GENERAR RETROALIMENTACIÓN/RESPUESTA
306
RESPUESTA VISUAL
307
RESPUESTA AUDITIVA
308

FIG. 3

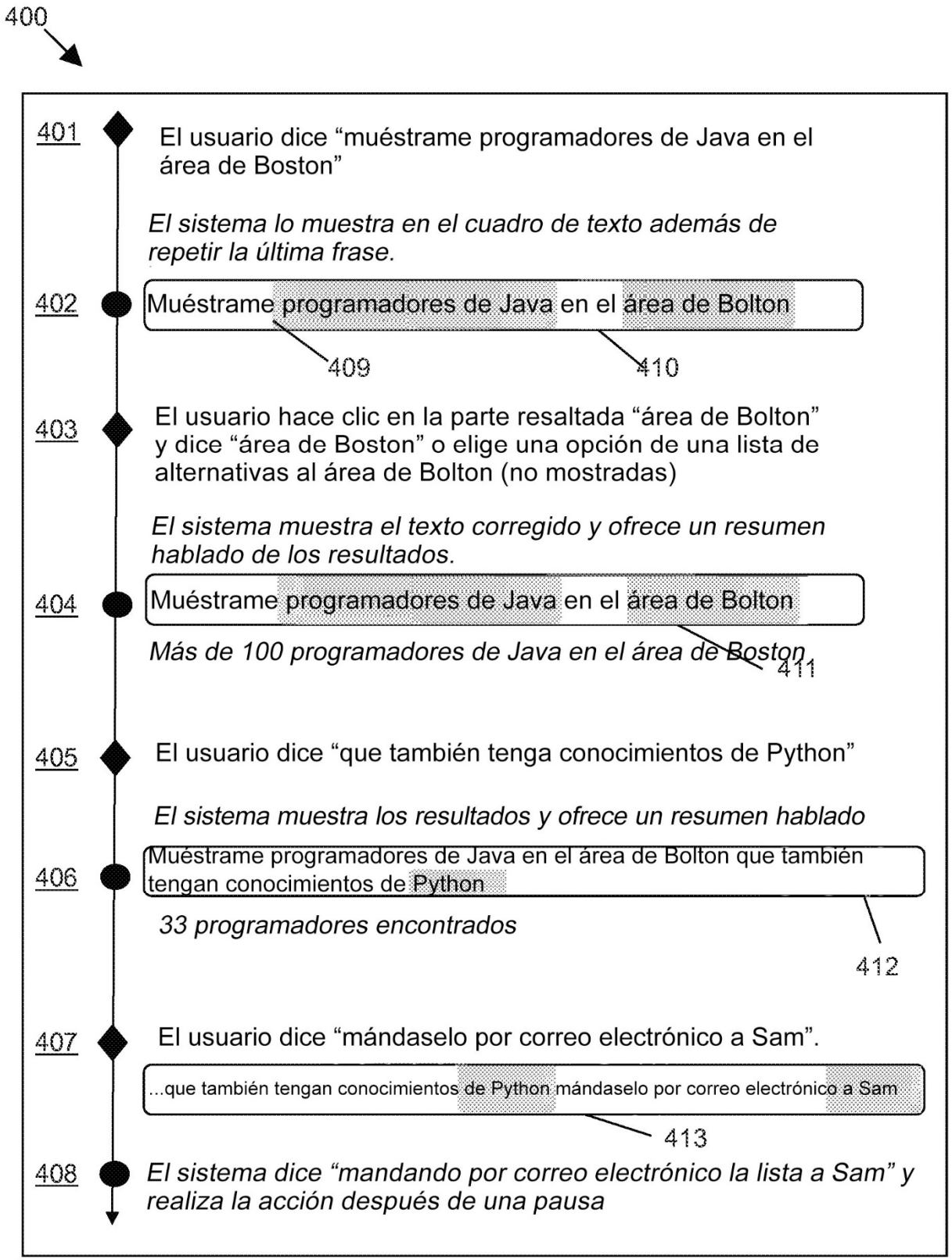
400
401
El usuario dice “muéstrame programadores de Java en el área de Boston”
El sistema lo muestra en el cuadro de texto además de repetir la última frase.
402
Muéstrame programadores de Java en el área de Bolton
409
410
403
El usuario hace clic en la parte resaltada “área de Bolton” y dice “área de Boston” o elige una opción de una lista de alternativas al área de Bolton (no mostradas)
El sistema muestra el texto corregido y ofrece un resumen hablado de los resultados.
404
Muéstrame programadores de Java en el área de Bolton
Más de 100 programadores de Java en el área de Boston
411
405
El usuario dice “que también tenga conocimientos de Python”
El sistema muestra los resultados y ofrece un resumen hablado
406
Muéstrame programadores de Java en el área de Bolton que también tengan conocimientos de Python
33 programadores encontrados
412
407
El usuario dice “mándaselo por correo electrónico a Sam”.
...que también tengan conocimientos de Python mándaselo por correo electrónico a Sam
413
408
El sistema dice “mandando por correo electrónico la lista a Sam” y realiza la acción después de una pausa

FIG. 4

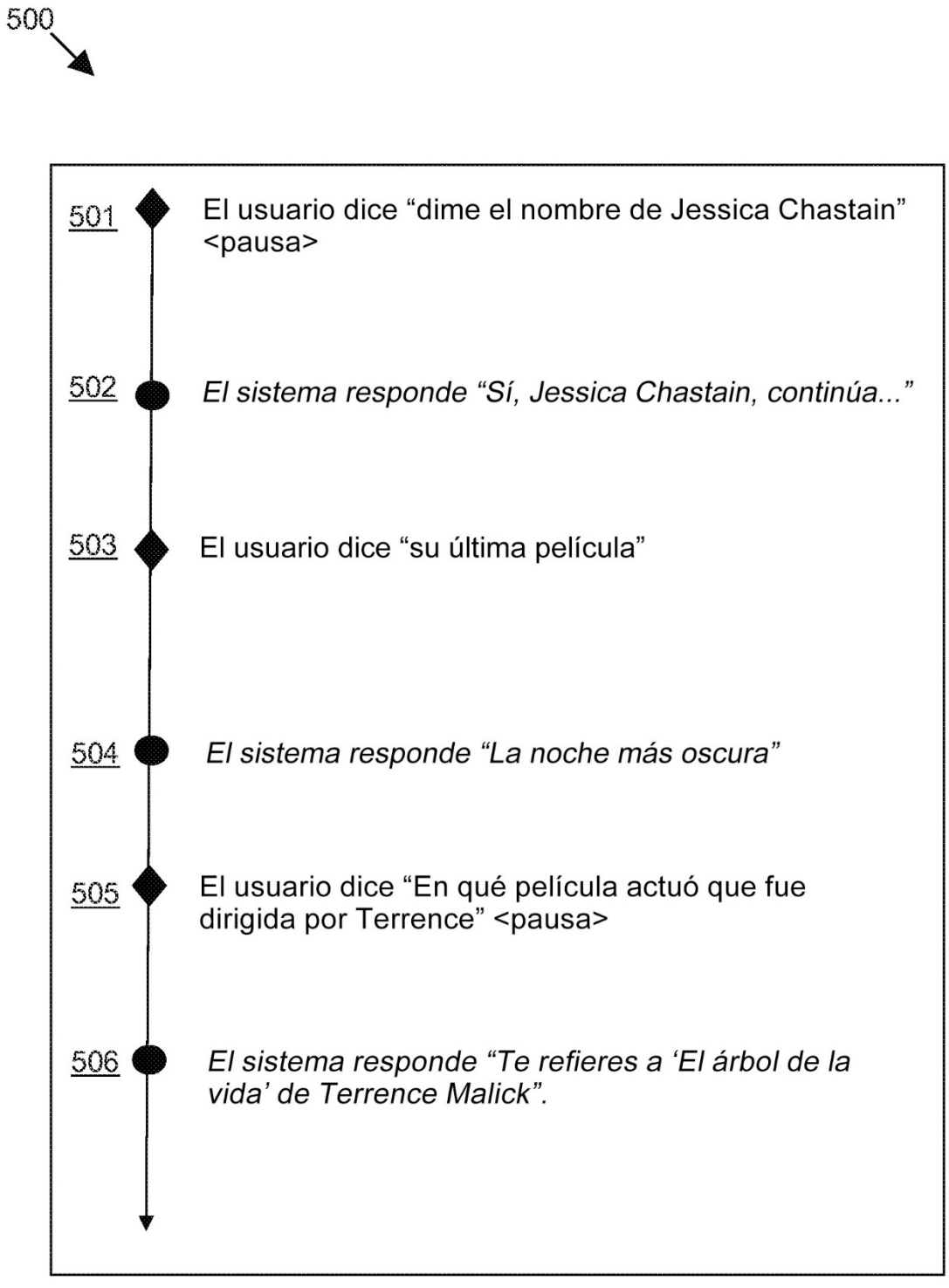
500
501
El usuario dice “dime el nombre de Jessica Chastain” <pausa>
502
El sistema responde “Sí, Jessica Chastain, continúa...”
503
El usuario dice “su última película”
504
El sistema responde “La noche más oscura”
505
El usuario dice “En qué película actuó que fue dirigida por Terrence” <pausa>
506
El sistema responde “Te refieres a ‘El árbol de la vida’ de Terrence Malick”.

FIG. 5

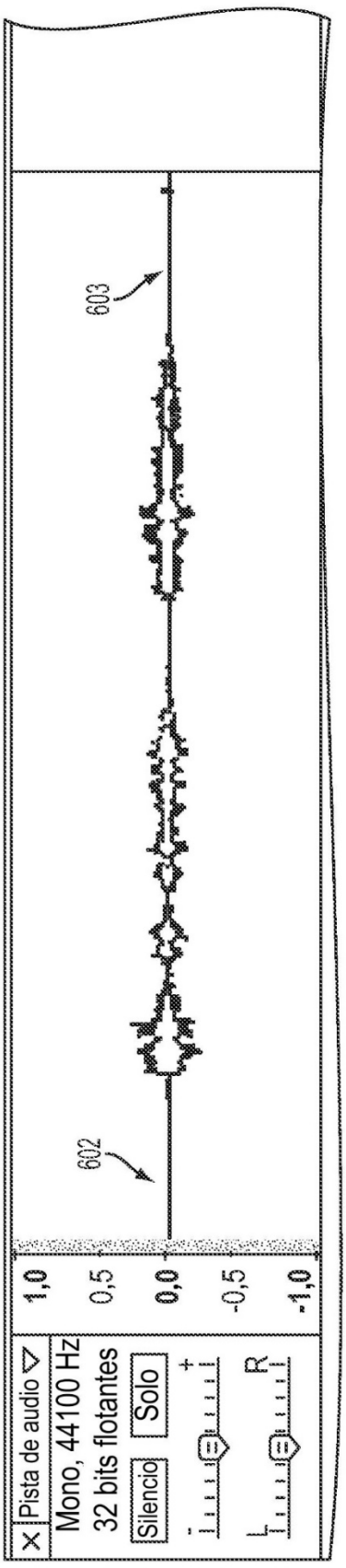
603
602
601
Pista de audio
Mono, 44100 Hz
32 bits flotantes
Silencio
Solo
1,0
0,5
0,0
-0,5
-1,0
R

FIG. 6A

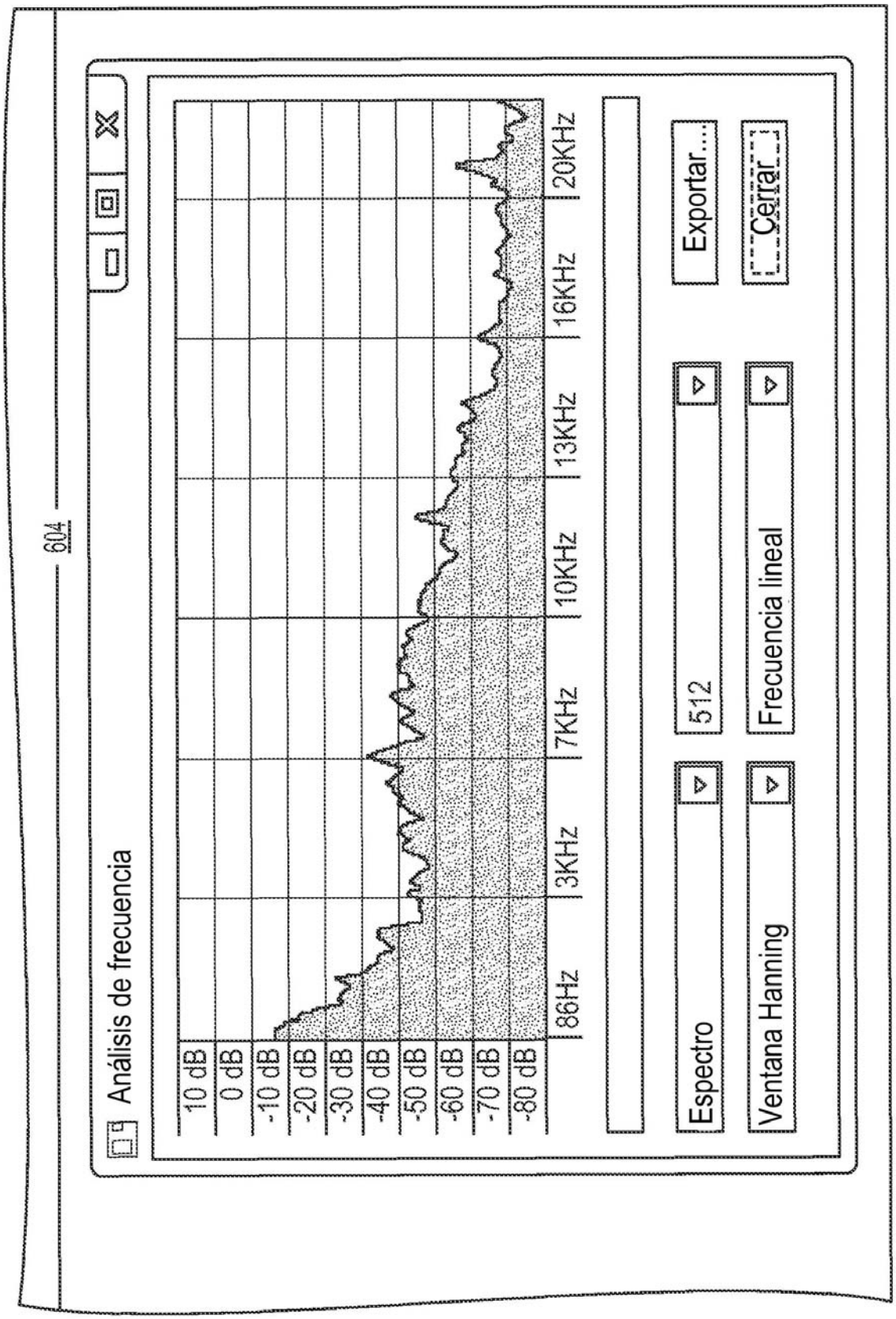
604
Análisis de frecuencia
10 dB
0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB
86Hz
3KHz
7KHz
10KHz
13KHz
16KHz
20KHz
Espectro
512
Exportar...
Ventana Hanning
Frecuencia lineal
Cerrar

FIG. 6B

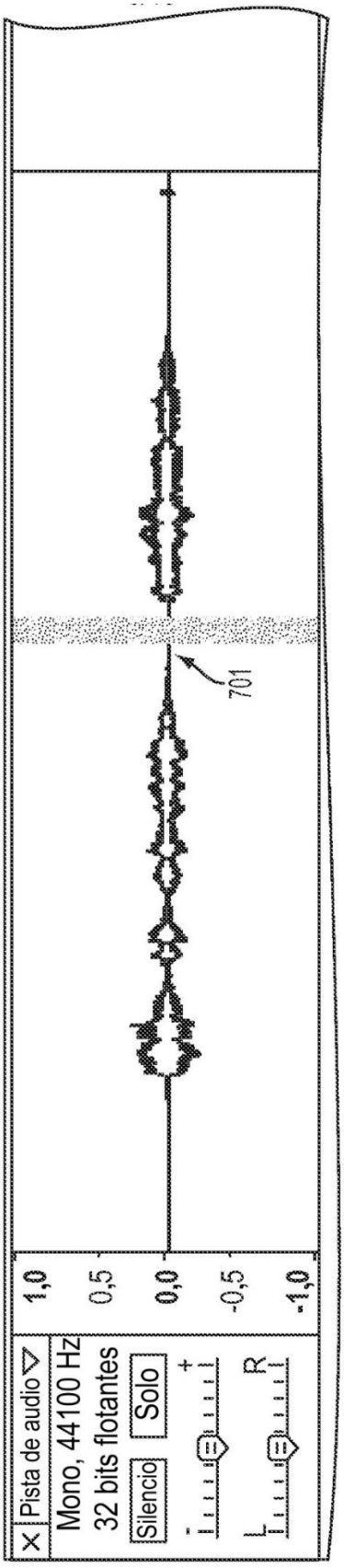
Pista de audio
Mono, 44100 Hz
32 bits flotantes
Silencio
Solo
L
R
1,0
0,5
0,0
-0,5
-1,0
701

FIG. 7A

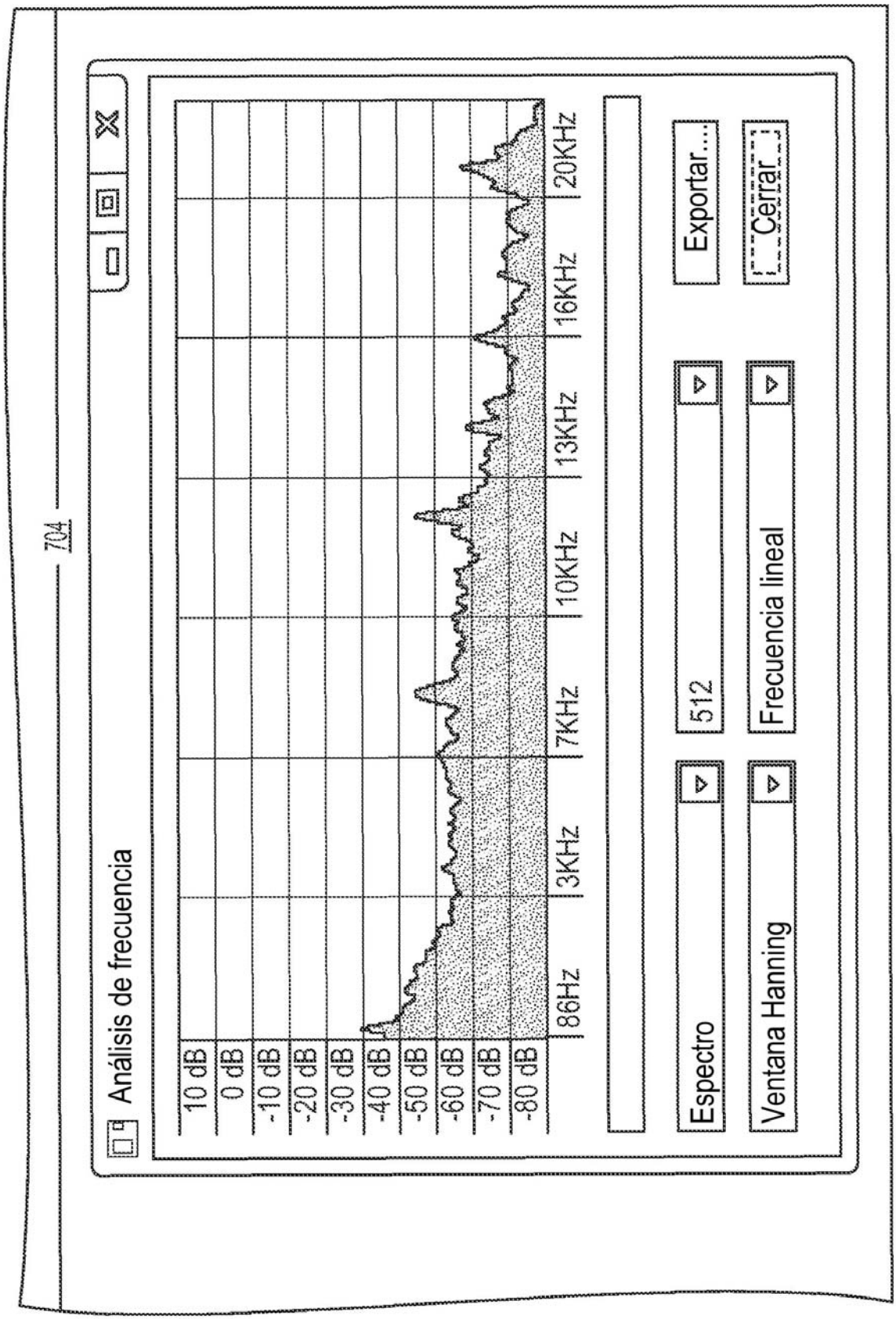
704
Análisis de frecuencia
10 dB
0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB
86Hz
3KHz
7KHz
10KHz
13KHz
16KHz
20KHz
Espectro
512
Exportar....
Ventana Hanning
Frecuencia lineal
Cerrar

FIG. 7B

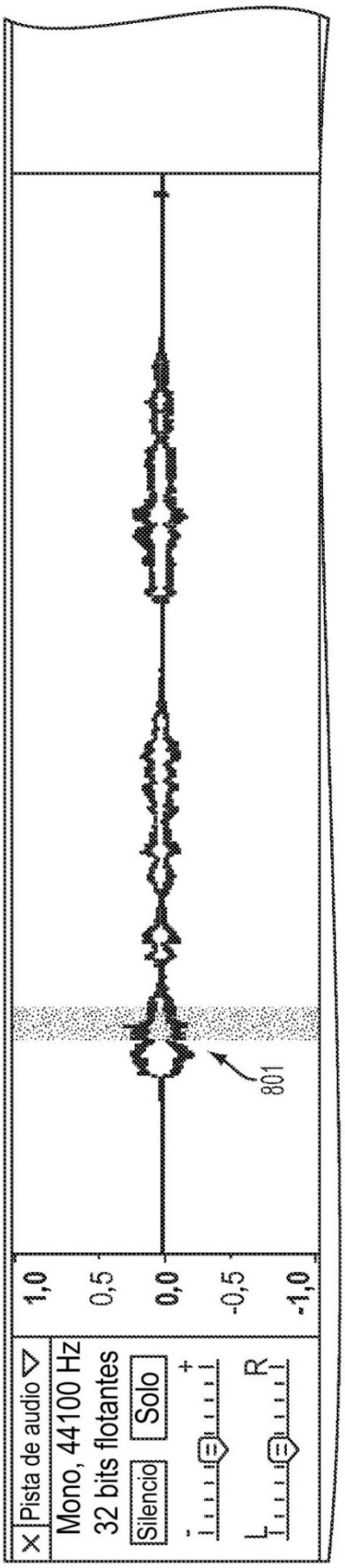
Pista de audio
Mono, 44100 Hz
32 bits flotantes
Silencio
Solo
L
R
1,0
0,5
0,0
-0,5
-1,0
801

FIG. 8A

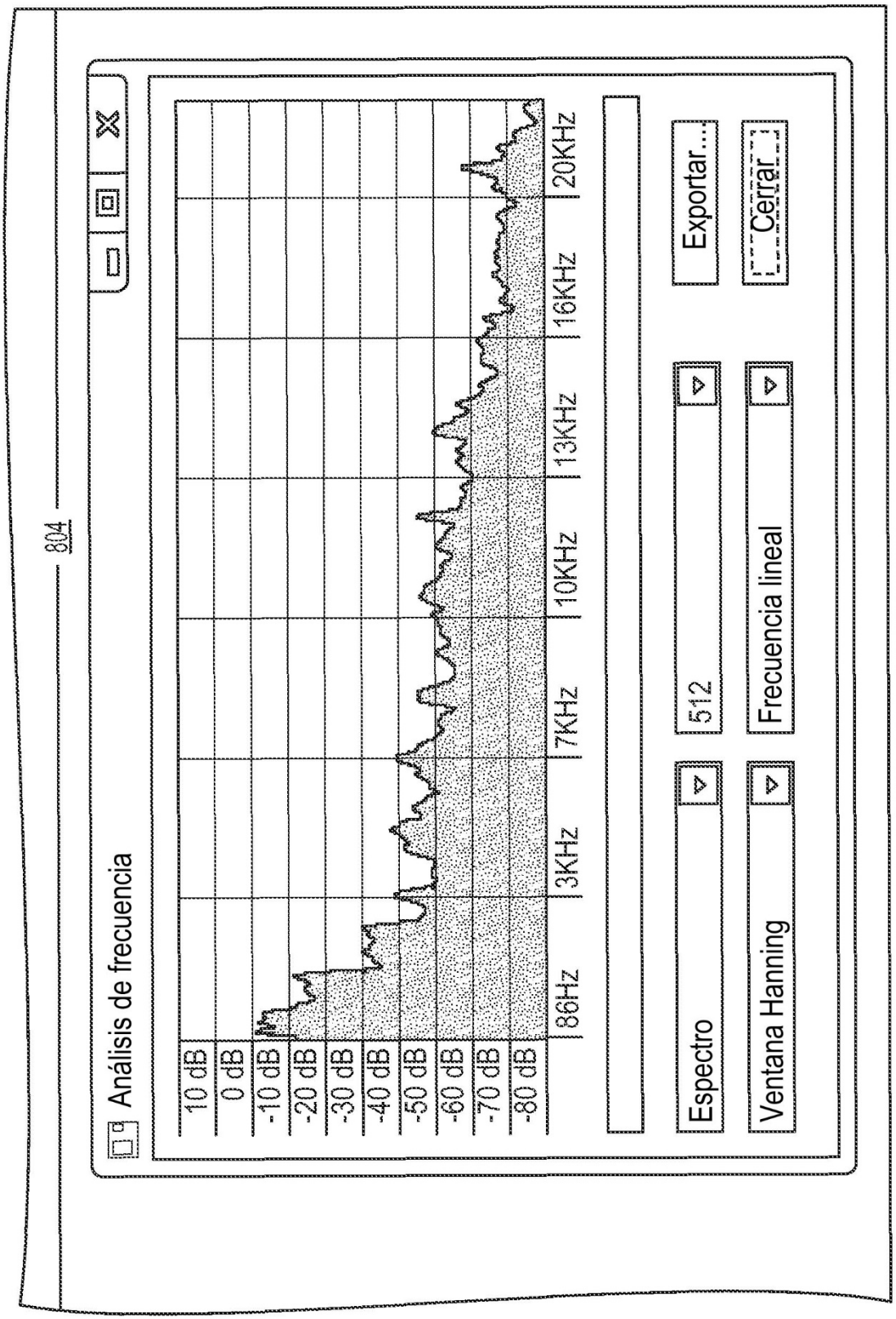
804
Análisis de frecuencia
10 dB
0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB
86Hz
3KHz
7KHz
10KHz
13KHz
16KHz
20KHz
Espectro
512
Exportar...
Ventana Hanning
Frecuencia lineal
Cerrar

FIG. 8B

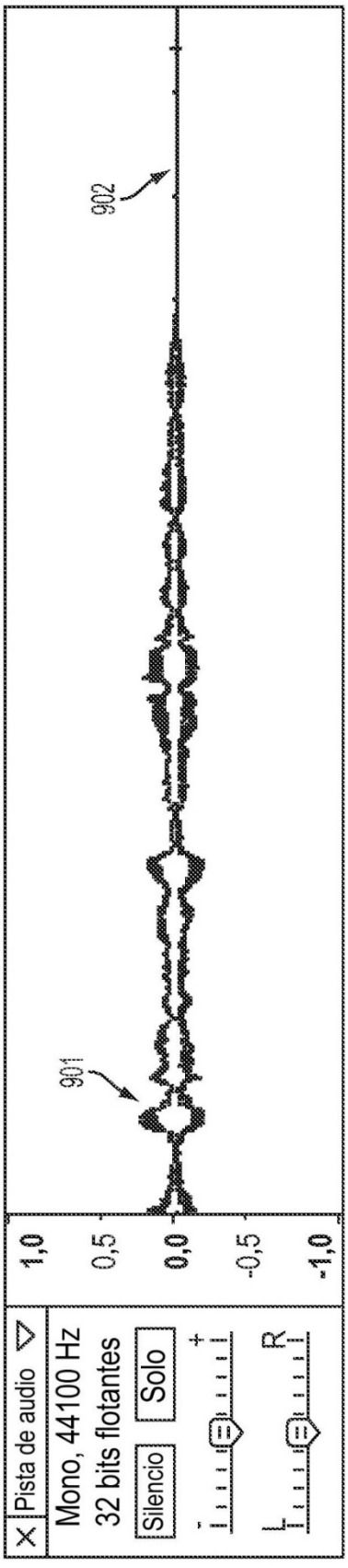
Pista de audio
Mono, 44100 Hz
32 bits flotantes
Silencio
Solo
1,0
0,5
0,0
-0,5
-1,0
901
902

FIG. 9

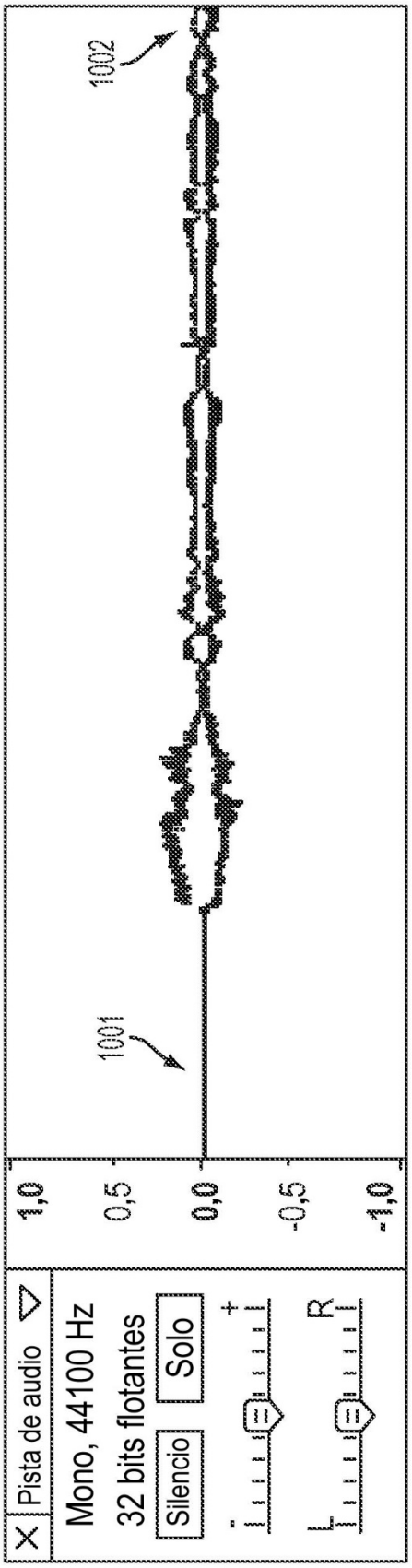
Pista de audio
Mono, 44100 Hz
32 bits flotantes
Silencio
Solo
L
R
1,0
0,5
0,0
-0,5
-1,0
1001
1002

FIG. 10