(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 446 667**

(51) Int. CI.:

G10L 15/30 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **31.08.2009 E 09810710 (5)**

(97) Fecha y número de publicación de la concesión europea: **13.11.2013 EP 2321821**

(54) Título: **Reconocimiento de voz distribuido utilizando una comunicación unidireccional**

(30) Prioridad:

29.08.2008 US 93221 P
30.08.2009 US 550381

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
10.03.2014

(73) Titular/es:

MULTIMODAL TECHNOLOGIES, LLC (100.0%)
1710 Murray Avenue
Pittsburgh, PA 15217, US

(72) Inventor/es:

CARRAUX, ERIC y
KOLL, DETLEF

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Reconocimiento de voz distribuido utilizando una comunicación unidireccional

ANTECEDENTES DE LA INVENCIÓN

Una diversidad de reconocedores de voz automáticos (ASRs) existen para realizar funciones tales como convertir voz en texto y controlar las operaciones de un ordenador en respuesta a la voz. Algunas aplicaciones de reconocedores de voz automáticos requieren tiempos de ida y retorno más cortos (la cantidad de tiempo entre cuando la voz es emitida y cuando el reconocedor de voz proporciona la salida) que otros para atender a las exigencias operativas del usuario final. A modo de ejemplo, un reconocedor de voz que se utiliza para una aplicación "en vivo" del reconocimiento de voz, tal como controlar el movimiento de un cursor en una pantalla, puede requerir un tiempo de ida y retorno más corto (también referido como un "tiempo de respuesta") que un reconocedor de voz que se utiliza para generar una transcripción de un informe médico.

El tiempo de respuesta deseado puede depender, a modo de ejemplo, del contenido de la vocalización que se procesa por el reconocedor de voz. A modo de ejemplo, para una vocalización de orden y control corta, tal como "cerrar ventana", un tiempo de respuesta superior a 500 ms puede parecer de lentitud excesiva para el usuario final. Al contrario, para una larga frase dictada que el usuario desea transcribir en texto, los tiempos de respuesta de 1000 ms pueden ser aceptables para el usuario final. De hecho, en este último caso, los usuarios pueden preferir tiempos de respuesta más largos porque, de no ser así, pueden sentir que su voz está siendo interrumpida por la visualización inmediata de texto en respuesta a su voz. Para dictados más largos, tales como párrafos enteros, tiempos de respuesta incluso más largos, de múltiples segundos pueden ser aceptables para el usuario final.

En los sistemas de reconocimiento de voz típicos de la técnica anterior, un tiempo de respuesta creciente mientras se mantiene la exactitud del reconocimiento requiere aumentar los recursos informáticos (ciclos de procesamiento y/o memoria) que se dediquen a realizar el reconocimiento de voz. En consecuencia, cualesquiera aplicaciones que requieran tiempos de respuesta rápidos exigen que el sistema de reconocimiento de voz se ejecute en el mismo ordenador en el que se ejecutan esas mismas aplicaciones. Aunque dicha ubicación puede eliminar el retardo que, de no ser así, se introduciría al exigir que los resultados del reconocimiento de voz sean transmitidos a la aplicación demandante a través de una red, dicha situación tiene también una diversidad de inconvenientes.

A modo de ejemplo, la ubicación operativa requiere que un sistema de reconocimiento de voz haya de instalarse en cada dispositivo del usuario final – tal como en cada ordenador de sobremesa, ordenador portátil, teléfono móvil y asistente digital personal (PDA) – lo que exige la funcionalidad del reconocimiento de voz. La instalación y mantenimiento de dichos sistemas de reconocimiento de voz en tan gran número y amplia diversidad de dispositivos puede ser tediosa y consumidora de tiempo para los usuarios finales y los administradores de sistemas. A modo de ejemplo, dicho mantenimiento requiere que se actualicen los denominados binarios de sistemas cuando una nueva versión del sistema de reconocimiento de voz se haga disponible. Los datos de los usuarios, tales como modelos de la voz, se crean y acumulan, en el transcurso del tiempo, sobre los dispositivos individuales, ocupando un espacio de almacenamiento de alto valor y la necesidad de sincronizarse con múltiples dispositivos utilizados por el mismo usuario. Dicho mantenimiento puede hacerse particularmente oneroso cuando los usuarios sigan utilizando los sistemas de reconocimiento de voz en un más amplio número y diversidad de dispositivos.

Además, la localización de un sistema de reconocimiento de voz en el dispositivo del usuario final hace que el sistema de reconocimiento de voz consuma recursos informáticos valiosos, tales como ciclos de procesamiento de la unidad central CPU, memoria principal y espacio de disco. Dichos recursos son particularmente escasos en dispositivos móviles portátiles, tales como teléfonos móviles. La producción de resultados del reconocimiento de voz, con tiempos de respuesta rápidos, utilizando tales dispositivos, exige normalmente menoscabar la exactitud del reconocimiento y reducir los recursos disponibles para otras aplicaciones que se ejecuten en el mismo dispositivo.

Una técnica conocida para superar estas limitaciones de recursos, en el contexto de dispositivos incorporados consiste en delegar parte o la totalidad de la responsabilidad del procesamiento del reconocimiento de voz a un servidor del reconocimiento de voz que esté situado a considerable distancia del dispositivo incorporado y que disponga de bastantes más recursos informáticos que el dispositivo incorporado. Cuando un usuario habla en el dispositivo incorporado en esta situación, el dispositivo incorporado no intenta reconocer la voz utilizando sus propios recursos operativos. En cambio, el dispositivo incorporado transmite la voz (o una forma procesada de la voz) por intermedio de una conexión de red al servidor de reconocimiento de voz, que reconoce la voz utilizando sus mayores recursos informáticos y, por lo tanto, produce resultados del reconocimiento con mayor rapidez que el dispositivo incorporado podría haber proporcionado con la misma exactitud. El servidor de reconocimiento de voz transmite, a continuación, los resultados de nuevo, a través de la conexión de la red, al dispositivo incorporado. En condiciones ideales, esta técnica produce resultados del reconocimiento de voz de gran exactitud y con mayor rapidez que sería posible, de no ser así, utilizando el dispositivo incorporado por sí solo.

En la práctica, sin embargo, esta técnica de "reconocimiento de voz en el lado del servidor" tiene numerosos inconvenientes. En particular, por cuanto que el reconocimiento de voz, en el lado del servidor, se basa en la

disponibilidad de conexiones de la red, de alta velocidad y fiables, la técnica falla si dichas conexiones no están disponibles cuando se necesiten. A modo de ejemplo, los potenciales aumentos en la velocidad, hechos posibles por el reconocimiento de voz en el lado del servidor, pueden negarse por el uso de una conexión de red sin un ancho de banda suficientemente alto. A modo de ejemplo, la latencia típica de la red de una llamada de HTTP a un servidor distante puede variar desde 100 ms a 500 ms. Si los datos hablados llegan a un servidor de reconocimiento de voz 500 ms después de emitirse, será imposible para ese servidor producir resultados con suficiente rapidez para actuar con el tiempo de respuesta mínimo (500 ms) requerido por las aplicaciones de órdenes y de control. En consecuencia, incluso el más rápido servidor de reconocimiento de voz producirá resultados que parezcan demasiado lentos si se usan en combinación con una conexión de red lenta.

Además, las técnicas convencionales de reconocimiento de voz, en el lado del servidor, suponen que la conexión de red establecida entre el cliente (p.e. dispositivo incorporado) y el servidor de reconocimiento de voz se mantiene activa continuamente durante el proceso de reconocimiento completo. Aunque hubiera la posibilidad de satisfacer esta condición en una Red de Área Local (LAN) o cuando ambos, cliente y servidor, sean gestionados por la misma entidad, esta condición puede ser imposible, o al menos irrazonable, de satisfacer cuando el cliente y el servidor estén conectados a través de una Red de Área Amplia (WAN) o de Internet, en cuyo caso, las interrupciones para la conexión de la red pueden ser frecuentes e inevitables.

Además, las organizaciones suelen restringir las clases de comunicaciones que sus usuarios puedan realizar a través de redes públicas tales como Internet. A modo de ejemplo, las organizaciones sólo pueden permitir a los clientes, dentro de sus redes, realizar comunicaciones salientes. Esto significa que un cliente puede entrar en contacto con un servidor externo en un determinado puerto, pero que el servidor no puede iniciar un contacto con el cliente. Lo que antecede es una realización, a modo de ejemplo, de una comunicación unidireccional.

Otra frecuente restricción impuesta sobre los clientes es que sólo pueden usar una gama limitada de puertos de salida para comunicarse con servidores externos. Además, puede exigirse que la comunicación saliente, en esos puertos, haya de estar encriptada. A modo de ejemplo, a los clientes se les suele permitir utilizar solamente el puerto HTTP estándar (puerto 80) o el puerto HTTPS estándar seguro encriptado (puerto 443).

Lo que se necesita, por lo tanto, son técnicas mejoradas para producir resultados de reconocimiento de voz, con tiempos de respuesta rápidos, sin sobrecargar los recursos informáticos limitados de los dispositivos del cliente. El documento US 7,330,815 describe un sistema de reconocimiento de voz, principalmente para el aprendizaje de idiomas, que comprende una pluralidad de clientes, que se pueden comunicar a través de Internet con un servidor que tenga servicios de reconocimiento de voz. El usuario cliente selecciona el localizador uniforme de recursos URL de un fichero en el servidor que soporte la técnica de reconocimiento de voz y un explorador en el cliente establece una conexión de TCP/IP a Internet y emite el URL utilizando esta conexión. El usuario demanda un fichero de HTML que comprenda una página web para uso en el procesamiento de la voz y esta página web será objeto de retorno al cliente. El explorador del cliente visualiza el contenido del fichero en la pantalla del cliente y el usuario puede seleccionar un ejercicio de procesamiento de la voz a partir del fichero visualizado. El Java script asociado con el ejercicio seleccionado activa un componente del explorador, que realiza el control del nivel de ventanas para capturar la voz desde el usuario. La voz se envía al servidor y el servidor envía debidamente una respuesta al cliente. El Java script establece un temporizador para uso en el sondeo operativo del componente del explorador para constatar si ha recibido una respuesta desde el servidor. La respuesta se transmite desde el servidor al Java script y a continuación, este último visualiza la respuesta al usuario. Un motor de reconocimiento de voz específico que puede utilizarse es el motor SAPI™ de Microsoft. Este motor procesa paquetes de voz desde el cliente, pero puede también estar en un periodo de temporización. Si el SAPI está en un periodo de temporización, la información que indica esta temporización es objeto de escritura en una memoria intermedia de salida y se hace retornar, como una respuesta de texto, al componente del explorador del cliente.

SUMARIO DE LA INVENCIÓN

Un cliente de reconocimiento de voz envía un flujo de voz y un flujo de control, en paralelo, a un reconocedor de voz, en el lado del servidor, a través de una red. La red puede ser una red de baja latencia, no operativamente fiable. El reconocedor de voz, en el lado del servidor, reconoce continuamente el flujo de voz. El cliente del reconocimiento de voz recibe los resultados del reconocimiento desde el reconocedor en el lado del servidor, en respuesta a las demandas desde el cliente. El cliente puede reconfigurar, a distancia, el estado del reconocedor, en el lado del servidor, durante el reconocimiento.

Otras características y ventajas de varios aspectos y formas de realización de la presente invención se harán evidentes a partir de la siguiente descripción y de las reivindicaciones.

La invención se establece por las reivindicaciones independientes 1, 6, 7 y 9.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La Figura 1 es un diagrama de flujo de datos de un sistema para realizar el reconocimiento de voz por intermedio de

una red de baja latencia según una forma de realización de la presente invención;

La Figura 2A es un diagrama de flujo de un método realizado por el sistema ilustrado en la Figura 1 según una forma de realización de la presente invención;

La Figura 2B es un diagrama de flujo de un método realizado por un reconocedor de voz automático, en el lado del servidor, para reconocer un segmento de voz según una forma de realización de la presente invención;

La Figura 2C es un diagrama de flujo de un método realizado por un reconocedor de voz automático, en el lado del servidor, como parte de la realización del reconocimiento de voz en segmentos de voz según una forma de realización de la presente invención;

La Figura 2D es un diagrama de flujo de un método realizado por un reconocedor, en el lado del servidor, para cerciorarse de que el reconocedor está reconfigurado después de que se hayan obtenido algunos resultados del reconocimiento y antes de que se realice un reconocimiento adicional según una forma de realización de la presente invención;

La Figura 3 es un diagrama de un flujo de voz según una forma de realización de la presente invención y

La Figura 4 es un diagrama de un flujo de órdenes y de control según una forma de realización de la presente invención.

DESCRIPCIÓN DETALLADA

Haciendo referencia a la Figura 1, se ilustra un diagrama de flujo de datos de un sistema de reconocimiento de voz 100 según una forma de realización de la presente invención. Con referencia a la Figura 2A, se ilustra un diagrama de flujo de un método 200 realizado por el sistema 100 de la Figura 1 según una forma de realización de la presente invención.

Un usuario 102 de un dispositivo cliente 106 habla y proporciona la voz 104 al dispositivo cliente 106 (etapa 202). El dispositivo cliente 106 puede ser cualquier dispositivo, tal como un ordenador de sobremesa u ordenador portátil, teléfono móvil, asistente digital personal digital (PDA), o teléfono fijo. Formas de realización de la presente invención, sin embargo, son de especial utilidad en conjunción con clientes de recursos limitados, tales como ordenadores o dispositivos informáticos móviles con procesadores lentos o pequeñas cantidades de memoria u ordenadores que ejecutan software con gran necesidad de recursos. El dispositivo 106 puede recibir la voz 104 desde el usuario 102 en cualquier forma, tal como por intermedio de un micrófono conectado a una tarjeta de sonido. La voz 104 puede materializarse en una señal de audio, que está tangiblemente memorizada en un medio legible por ordenador y/o transmitido a través de una conexión de red u otro canal. La voz 104 puede, a modo de ejemplo, incluir múltiples flujos de audio, como en el caso de aplicaciones de tipo "*push to talk*" (‘pulsar para hablar’), en las que cada pulsación inicia un nuevo flujo de audio.

El dispositivo cliente 106 incluye una aplicación 108, tal como una aplicación de transcripción u otra aplicación que necesite reconocer la voz 104. Aunque la aplicación 108 puede ser cualquier clase de aplicación que utiliza los resultados del reconocimiento de voz, se supone para los fines de la siguiente descripción que la aplicación 108 es una aplicación del reconocimiento “en vivo” para transcribir la voz. Partes de la voz 104, proporcionada por el usuario 102 en este contexto, puede caer dentro de una de dos categorías básicas: voz dictada a transcribirse (p.e., "El paciente es un hombre de 35 años de edad") u órdenes (tales como "suprima esto" o "firme y presente").

El dispositivo del cliente 106 incluye también un cliente del reconocimiento de voz 140. Aunque el cliente del reconocimiento de voz 140 se ilustra en la Figura 1 como un módulo separado de la aplicación 108, como alternativa, el cliente del reconocimiento de voz 140 puede ser parte de la aplicación 108. La aplicación 108 proporciona la voz 104 al cliente de reconocimiento de voz 140. Como alternativa, la aplicación 108 puede procesar la voz 104 en alguna forma y proporcionar la versión procesada de la voz 104, u otros datos derivados a partir de la voz, al cliente del reconocimiento de voz 140. El cliente del reconocimiento de voz 140, por sí mismo, puede procesar la voz 104 (además o en lugar de cualquier procesamiento realizado sobre la voz por la aplicación 108) en preparación para transmitir la voz 104 para su reconocimiento.

El cliente del reconocimiento de voz 140 transmite la voz 104 a través de una red 116 a un motor de reconocimiento de voz 120, en el lado del servidor, situado en un servidor de reconocimiento de voz 118 (etapa 204). Aunque el cliente 140 puede transmitir la voz completa 104 al servidor 118 utilizando una configuración de servidor única, hacerlo así puede producir resultados no óptimos. Para mejorar la exactitud del reconocimiento o cambiar el contexto del motor de reconocimiento de voz 120, el cliente 140 puede, en cambio, reconfigurar el motor de reconocimiento de voz 120 en varios puntos durante la transmisión de la voz 104 y por lo tanto, en varios puntos durante el reconocimiento de la voz del motor de reconocimiento de voz 104. En general, las órdenes de configuración transmitidas por el cliente 140 al motor de reconocimiento de voz 120 establecen las expectativas del reconocedor 120 con respecto al contexto y/o contenido de la voz que ha de seguir. Varios sistemas de la técnica

anterior realizan esta función de configuración configurando el motor de reconocimiento, en el lado del servidor, con una configuración inicial, enviando luego parte de la voz al servidor, reconfigurando, a continuación, el motor de reconocimiento en el lado del servidor, y enviando luego más de la voz y así sucesivamente. Esto permite al motor de reconocimiento, en el lado del servidor, reconocer diferentes partes de la voz con configuraciones y en contextos que están diseñados para producir mejores resultados para partes posteriores de la voz que se habrían producido utilizando la configuración inicial.

Es indeseable, sin embargo, requerir al cliente del reconocimiento de voz 140 que espere a recibir una confirmación desde el servidor 118 de que la orden de reconfiguración anterior se ha procesado por el servidor 118 antes de enviar la siguiente parte de la voz 104 al servidor 118, porque dicho requisito podría introducir un retardo importante en el reconocimiento de la voz 104, en particular si la conexión de la red es lenta y/o no fiable. También es indeseable interrumpir el procesamiento de la voz, en el lado del servidor, hasta que el servidor reciba instrucciones desde la aplicación 108, en el lado del cliente, sobre cómo procesar la voz subsiguiente. En los sistemas de la técnica anterior, sin embargo, el servidor necesita interrumpir el procesamiento de la voz hasta que reciba dichas instrucciones, tales como nuevas órdenes de configuración, desde el cliente.

Las formas de realización de la presente invención resuelven estos y otros problemas como sigue. El cliente del reconocimiento de voz 140 transmite la voz 104 al servidor 118 en un flujo de voz 110 a través de la red 116 (FIG. 2, etapa 204). Según se ilustra en la Figura 3, el flujo de voz 110 puede dividirse en segmentos 302a-e, pudiendo cada segmento representar una parte de la voz 104 (p.e.150-250 ms de la voz 104). El envío de la voz 104 en segmentos permite al cliente del reconocimiento de voz 140 transmitir partes de la voz 104 al servidor 118 relativamente pronto después de que esas partes se hagan disponibles para el cliente del reconocimiento de voz 140, permitiendo así al reconocedor 120 comenzar a reconocer dichas partes con el mínimo retardo. La aplicación 108 puede, a modo de ejemplo, enviar el primer segmento 302a inmediatamente después de que se haga disponible, incluso cuando se está generando el segundo segmento 302b. Además, el cliente 140 puede transmitir partes individuales en el flujo de voz 110 al servidor 118 sin utilizar una conexión estacionaria (p.e. toma de corriente). En consecuencia, un protocolo sin conexión o sin estados, tal como HTTP, puede usarse por el cliente del reconocimiento de voz 140 para transmitir el flujo de voz 110 al servidor 118.

Aunque solamente cinco segmentos representativos 302a-e se ilustran en la Figura 2A para facilidad ilustración, en la práctica, el flujo de voz 110 puede contener cualquier número de segmentos, que puede aumentar mientras el usuario 102 sigue hablando. La aplicación 108 puede usar cualquier procedimiento para dividir la voz 104 en segmentos, o para dirigir un flujo de la voz 104 hacia el servidor 118, A modo de ejemplo, una conexión de HTTP.

Cada uno de los segmentos de voz 302a-e contiene datos 304a que representan una parte correspondiente de la voz 104 del usuario 102. Dichos datos de voz 304a pueden representarse en cualquier formato adecuado. Cada uno de los segmentos de la voz 302a-e puede contener otra información, tal como el tiempo de inicio 304b y el tiempo final 304c de los datos de la voz 304a correspondientes y una etiqueta operativa 304d que se describirá, a continuación, con más detalle. Los campos particulares 304a-d ilustrados en la Figura 3 son simplemente a modo de ejemplo y no constituyen limitaciones de la presente invención.

En general, el reconocedor en el lado del servidor 120 pone en cola de espera a segmentos desde el flujo de voz 110 en la cola de espera de procesamiento 124, del tipo de primero en entrar primero en salir, en el servidor 118 (Figura 2, etapa 216). Con algunas excepciones que se describirán, a continuación, con más detalle, el reconocedor en el lado del servidor 120 extrae segmentos desde la cola de espera de procesamiento 124, tan pronto como sea posible, después de que se hagan disponibles y realiza el reconocimiento de voz en esos segmentos para producir resultados de reconocimiento de voz (etapa 218), cuyo servidor 120 los coloca en una cola de espera de salida 134 (etapa 220) del tipo de primero en entrar primero en salir.

La aplicación 108, por intermedio del cliente del reconocimiento de voz 140, puede enviar también un flujo de control 112 al reconocedor en el lado del servidor 120 a través de la red 116 como parte de la etapa 204. Según se ilustra en la Figura 4, el flujo de control 112 puede incluir mensajes de control 402a-c, transmitidos en secuencia al reconocedor 120. Aunque solamente tres mensajes de control representativo 402a-c se ilustran en la Figura 4 para facilidad de ilustración, en la práctica el flujo de control 112 puede contener cualquier número de mensajes de control. Como se describirá con más detalle a continuación, cada uno de los mensajes de control 402a puede contener una pluralidad de campos, tal como un campo de órdenes 404a para especificar una orden que ha de ejecutarse por el reconocedor en el lado del servidor 120, un campo de objetos de configuración 404b para especificar un objeto de configuración y un campo del valor de temporización 404c para especificar un valor de temporización. Los campos particulares 304a-d ilustrados en la Figura 3 son simplemente ejemplos y no constituyen limitaciones de la presente invención.

Según se ilustra en la Figura 1, el cliente del reconocimiento de voz 140 puede tratar el flujo de voz 110 y el flujo de control 112 como dos flujos de datos diferentes (etapas 206 y 208), transmitidos en paralelo desde el cliente del reconocimiento de voz 140 al motor 120. Sin embargo, suponiendo que solamente un puerto de salida está disponible para el cliente del reconocimiento de voz 140 para su comunicación con el servidor 118, el cliente 106 puede multiplexar el flujo de voz 110 y el flujo de control 112 en un flujo de datos único 114 transmitido al servidor

118 (etapa 210). El servidor 118 demultiplexa la señal 114 en su flujo de voz constituyente 110 y el flujo de control 112 en el lado del servidor (etapa 214).

Se puede utilizar cualquier sistema de multiplexación. A modo de ejemplo, si HTTP se usa como un mecanismo de transporte en tal caso, un cliente de HTTP 130 y un servidor de HTTP 132 pueden realizar, de forma transparente, las funciones de multiplexación y de demultiplexación, respectivamente, en las funciones del cliente 106 y del servidor 118. Dicho de otro modo, el cliente del reconocimiento de voz 140 puede tratar el flujo de voz 110 y el flujo de control 112 como dos flujos separados aun cuando se transmitan como un flujo multiplexado único 114 porque el cliente de HTTP 130 multiplexa estos dos flujos juntos, de forma automática y transparente, en las funciones del cliente del reconocimiento de voz 140. De modo similar, el reconocedor en el lado del servidor 120 puede tratar el flujo de voz 110 y el flujo de control 112 como dos flujos separados aun cuando se reciban por el servidor 118 como un flujo multiplexado único 114 porque el servidor de HTTP 132 demultiplexa el flujo combinado 114 en dos flujos, de forma automática y transparente, en la función del reconocedor en el lado del servidor 120.

Según se indicó con anterioridad, por defecto, el reconocedor en el lado del servidor 120 extrae segmentos de la voz desde la cola de espera de procesamiento 124 en secuencia, realiza el reconocimiento de voz en ellos y pone en cola de espera a los resultados del reconocimiento de voz en la cola de espera de salida 134. El cliente del reconocimiento de voz 108 recibe los resultados del reconocimiento de voz como sigue. El cliente del reconocimiento de voz 140 envía, en el flujo de control 112, un mensaje de control cuyo campo de órdenes 404a llama un método aquí referido como "DecodeNext." Este método toma como parámetros un objeto de actualización de configuración 404b (que especifica cómo ha de actualizarse un estado de configuración 126 del reconocedor en el lado del servidor 120), y un valor de temporización 404c en tiempo real. Aunque el cliente del reconocimiento de voz 140 puede enviar otras órdenes en el flujo de control 112, solamente la orden de DecodeNext se describirá aquí para mayor facilidad de explicación.

El reconocedor en el lado del servidor 120 extrae mensajes de control desde el flujo de control 112 en secuencia, tan pronto como sea posible después de que se reciban, y en paralelo con el procesamiento de los segmentos de la voz en el flujo de voz 110 (etapa 222). El reconocedor en el lado del servidor 120 ejecuta la orden en cada mensaje de control en secuencia (etapa 224).

Haciendo referencia a la Figura 2B, se ilustra un diagrama de flujo de un método realizado por el reconocedor en el lado del servidor 120 para ejecutar un mensaje de control de DecodeNext en el flujo de control 112. Si al menos un resultado del reconocimiento de voz está en la cola de espera de salida 134 (etapa 240), el reconocedor 120 envía los siguientes resultados 122 en la cola de espera 134 al cliente del reconocimiento de voz 140 a través de la red 116 (etapa 242). Si más de un resultado está disponible en la cola de espera 134 en el momento en que se realiza la etapa 242, entonces todos los resultados disponibles en la cola de espera 134 se transmiten en el flujo de resultados 122 al cliente del reconocimiento de voz 140. (Aunque los resultados 122 se ilustren en la Figura 1 como siendo transmitidos directamente desde el reconocedor 120 al cliente del reconocimiento de voz 140 para facilidad de ilustración, los resultados 122 pueden transmitirse por el servidor de HTTP 132 a través de la red 116 y recibirse por el cliente de HTTP 130 en el dispositivo cliente 106.). El método de DecodeNext devuelve, entonces, el control a la aplicación 108 (etapa 246) y termina el proceso.

No hay que olvidar que el reconocedor 120 está realizando continuamente el reconocimiento de voz en los segmentos de voz en la cola de espera de procesamiento 124. Por lo tanto, si la cola de espera de salida 134 está vacía cuando el reconocedor 120 comienza a ejecutar el método de DecodeNext, el método de DecodeNext se bloquea hasta que al menos un resultado (p.e. una palabra) esté disponible en la cola de espera de salida 134, o hasta que se alcance la cantidad de tiempo especificada por el valor de temporización 404c (etapa 248). Si un resultado aparece en la cola de espera de salida 134 antes de que se alcance el valor de temporización 404c, entonces el método de DecodeNext transmite ese resultado al cliente del reconocimiento de voz 140 (etapa 242), devuelve el control al cliente del reconocimiento de voz 140 (etapa 246), y termina el proceso. Si ningún resultado aparece en la cola de espera de salida 134 antes de que se alcance el valor de temporización 404c, entonces, el método de DecodeNext informa al cliente del reconocimiento de voz 140 de que ningún resultado está disponible (etapa 244), devuelve el control al cliente del reconocimiento de voz 140 (etapa 246) y termina el proceso sin reenviar cualesquiera resultados del reconocimiento al cliente de reconocimiento de voz 140.

Una vez que vuelve el control al cliente del reconocimiento de voz 140 (después de que el método de DecodeNext reenvíe un resultado del reconocimiento al cliente del reconocimiento de voz 140 o informa al cliente del reconocimiento de voz 140 de que ninguno de dichos resultados está disponible), el cliente del reconocimiento de voz 140 puede enviar inmediatamente otro mensaje de DecodeNext al servidor 120 en un intento para recibir el siguiente resultado del reconocimiento. El servidor 120 puede procesar este mensaje de DecodeNext en la manera antes descrita con respecto a la Figura 2B. Este proceso se puede repetir para posteriores resultados del reconocimiento. En consecuencia, el flujo de control 112 puede estar esencialmente siempre en estado de bloqueo en el lado del servidor (en el bucle representado por las etapas 240 y 248 en la Figura 2B), a la espera de resultados del reconocimiento y reenviándolos a la aplicación del cliente 108 cuando se hicieren disponibles.

El valor de temporización 404c puede elegirse para ser más corto que el valor de temporización del protocolo de

comunicación subyacente usado entre el cliente 140 y el servidor 120, tal como el valor de temporización de HTTP En consecuencia, si el cliente 140 recibe la notificación desde el servidor de que ningún resultado del reconocimiento de voz se produjo antes de que fuera alcanzado el valor de temporización 404c, el cliente 140 puede llegar a la conclusión de que la temporización de espera era el resultado de la incapacidad del servidor 120 para producir cualesquiera resultados del reconocimiento de voz antes de que fuera alcanzado el valor de temporización 404c, en lugar de como el resultado de un problema de comunicación de la red. Sea cual fuere el motivo para la temporización, sin embargo, el cliente 140 puede enviar otro mensaje DecodeNext al servidor 120 después de dicho periodo de temporización.

Las formas de realización, a modo de ejemplo, anteriores se refieren a dos flujos de datos 110 y 112 completamente no sincronizados. Sin embargo, puede ser deseable realizar algunas clases de sincronización en los dos flujos 110 y 112. A modo de ejemplo, puede ser de utilidad para el cliente del reconocimiento de voz 140 cerciorarse de que el reconocedor 120 está en un determinado estado de configuración antes de comenzar a reconocer el flujo de voz 110. A modo de ejemplo, el reconocedor 120 puede utilizar el contexto textual de la posición del cursor actual en una ventana de edición de texto para guiar el reconocimiento para texto que haya de insertarse en esa posición del cursor. Puesto que la posición del cursor puede cambiar con frecuencia debido a la acción del ratón o de otra pulsación del teclado, puede ser de utilidad para la aplicación 108 retardar la transmisión del contexto del texto para el servidor 120 hasta que el usuario 102 pulse el botón "start recording" (‘inicio del registro’). En este caso, debe impedirse al reconocedor en el lado del servidor 120 reconocer la voz transmitida al servidor 120 hasta que se reciba el contexto del texto por el servidor 120 y el servidor 120 actualice su estado de configuración 126 en consecuencia.

A modo de otro ejemplo, algunos resultados del reconocimiento pueden iniciar la necesidad de cambiar el estado de configuración 126 del reconocedor 120. En consecuencia, cuando el reconocedor en el lado del servidor 120 genere dicho resultado, debe esperar hasta que se reconfigure antes de generar el siguiente resultado. A modo de ejemplo, si el reconocedor 120 produces el resultado de "suprimir todo", la aplicación 108 puede intentar, a continuación, verificar la intención del usuario preguntando al usuario 102 lo que sigue: "¿Desea realmente suprimir todo? Diga SÍ o NO". En este caso, la aplicación 108 (por intermedio del cliente del reconocimiento de voz 140) debe reconfigurar el reconocedor 120 con una respuesta "SÍ I NO" antes de que el reconocedor 120 intente reconocer el siguiente segmento en el flujo de voz 110.

Dichos resultados pueden obtenerse como sigue, según se indica por el diagrama de flujo de la Figura 2C, que ilustra un método que puede realizarse por el reconocedor en el lado del servidor 120 colmo parte de la realización del reconocimiento de voz en los segmentos de audio en la cola de espera del procesamiento (Figura 2A, etapa 218). A cada estado de configuración del reconocedor se le asigna un identificador único del estado de configuración (ID). El cliente del reconocimiento de voz 140 asigna valores enteros a los identificadores IDs del estado de configuración, de modo que si ID1 > ID2, entonces el estado de configuración asociado con ID1 es más reciente que el estado de configuración asociado con ID2. Según se describió antes con respecto a la Figura 3, el cliente del reconocimiento de voz 140 proporciona también etiquetas operativas 304d dentro de cada uno de los segmentos del flujo de voz 302a-e que indican el número mínimo del ID del estado de configuración ID que se requiere antes de que se pueda iniciar el reconocimiento de ese segmento.

Cuando el reconocedor en el lado del servidor 120 recupera el siguiente segmento de audio desde la cola de espera del procesamiento 124 (etapa 262), el reconocedor 120 compara el ID del estado de configuración 136 del estado de configuración 126 actual del reconocedor con el ID de configuración mínimo requerido, que se especifica por la etiqueta operativa del segmento de audio 304d objeto de recuperación. Si el ID de configuración 136 actual es al menos tan grande como el ID de configuración mínimo requerido (etapa 264), entonces, el servidor 120 inicia el reconocimiento del segmento de audio recuperado (etapa 266). De no ser así, el servidor 120 queda a la espera hasta que su ID de configuración 136 alcance el ID mínimo requerido antes de que inicie el reconocimiento del segmento de voz actual. Puesto que el método de la Figura 2C puede realizarse en paralelo con el método 200 representado en la Figura 2A, el ID de configuración 136 del reconocedor en el lado del servidor 120 puede actualizarse mediante la ejecución de mensajes de control 224 incluso mientras el método de la Figura 2C bloquea en el bucle en la etapa 264. Además, conviene señalar que incluso cuando el servidor 120 está a la espera de procesar la voz desde la cola de espera del procesamiento 124, el servidor 120 sigue recibiendo segmentos adicionales desde el flujo de voz 110 y coloca esos segmentos en la cola de espera del procesamiento 124 (Figura 2A, etapas 214, 216).

A modo de otro ejemplo de formas en las que el flujo de voz 110 y el flujo de control 112 pueden sincronizarse, la aplicación 108, por intermedio del cliente del reconocimiento de voz 140, puede dar instrucciones al reconocedor 120 con anticipación a la interrupción del reconocimiento del flujo de voz 110, o tomar alguna otra acción, al producir cualquier resultado del reconocimiento o al producir un resultado del reconocimiento que satisfaga algunos criterios. Dichos criterios pueden servir efectivamente como puntos de ruptura operativa en donde la aplicación 108, por intermedio de del cliente del reconocimiento de voz 140, puede utilizarse para un control proactivo de la magnitud de la anticipación con la que el reconocedor 120 produce resultados del reconocimiento.

A modo de ejemplo, se considera un contexto en el que el usuario 102 puede emitir cualquiera de las órdenes de voz siguientes: "suprimir", "siguiente", "seleccionar todo" y "abrir seleccionador de fichero". En este contexto, una

posible configuración, que puede especificarse por el objeto de actualizar la configuración 404b, sería: <suprimir, continuar>, <siguiente, continuar>, <seleccionar todo, continuar>, <abrir seleccionador fichero, parar>. Dicha configuración da instrucciones al reconocedor en el lado del servidor 120 para continuar reconociendo el flujo de voz 110 después de obtener el resultado del reconocimiento "suprimir", "siguiente" o "seleccionar todo," pero interrumpir el reconocimiento del flujo de voz 110 después de obtener el resultado del reconocimiento " abrir seleccionador fichero ". El motivo para configurar el reconocedor 120, de esta forma, es que la producción de los resultados "suprimir", "siguiente" o "seleccionar todo" no requiere que el reconocedor 120 haya de reconfigurarse antes de producir el siguiente resultado. Por lo tanto, al reconocedor 120 se le puede permitir que prosiga reconociendo el flujo de voz 110 después de proporcionar cualquiera de los resultados de "suprimir", "siguiente" o "seleccionar todo", lo que permite al reconocedor 120 seguir el reconocimiento de la voz 104 a plena velocidad (ver Figura 2D,etapa 272). Por el contrario, la producción del resultado " abrir seleccionador fichero" requiere al reconocedor 120 que ha de reconfigurarse (p.e., para esperar resultados previstos tales como "OK," "seleccionar file1.xml," o "Nueva Carpeta") antes de reconocer cualesquiera segmentos subsiguientes en el flujo de voz 110 (ver Figura 2C, etapa 274). Por lo tanto, si la aplicación 108, por intermedio del cliente del reconocimiento de voz 140, es informada por el reconocedor 120 de que se produjo el resultado " abrir seleccionador fichero", la aplicación 108, por intermedio del cliente del reconocimiento de voz 140, puede reconfigurar el reconocedor 120 con un estado de configuración que sea adecuado para el control de un seleccionar de fichero. Al permitir a la aplicación 108 preconfigurar el reconocedor 120, de esta forma, plantea un equilibrio entre maximizar el tiempo de respuesta del reconocedor y asegurar que el reconocedor 120 utilice el estado de configuración adecuado para reconocer diferentes partes de la voz 104.

Conviene señalar que aun cuando el reconocedor 120 interrumpa el reconocimiento de la voz desde la cola de espera de procesamiento 124 como el resultado de una orden de “parada” de la configuración (etapa 274), el reconocedor 120 puede seguir recibiendo segmentos de voz desde el flujo de voz 110 y poner a esos segmentos en la cola de espera del procesamiento 124 (Figura 2A, etapas 214, 216). En consecuencia, segmentos adicionales del flujo de voz 110 están preparados para procesarse tan pronto como el reconocedor 120 reanude la realización del reconocimiento de voz.

Según se indicó con anterioridad, las técnicas aquí dadas a conocer pueden usarse en conjunción con protocolos de comunicaciones unidireccionales, tales como HTTPS. Dichos protocolos de comunicaciones son sencillos de configurar en redes de área amplia, pero ofrecen poca garantía contra los fallos operativos. Los fallos pueden ocurrir durante una demanda entre el cliente 130 y el servidor 132 que puede dejar a la aplicación 108 en un estado ambiguo. A modo de ejemplo, un problema puede ocurrir cuando una u otra parte (aplicación del cliente 108 o reconocedor en el lado del servidor 120) falla mientras en el curso de una llamada. Otros problemas puede ocurrir, a modo de ejemplo, debido a mensajes perdidos hacia o desde el servidor 118, mensajes que llegan al cliente 106 o al servidor 118 fuera de secuencia o mensajes enviados, por error, colmo duplicados. En general, en los sistemas de la técnica anterior, es la responsabilidad del cliente del reconocimiento de voz 140 asegurar la solidez operativa del sistema global 100, puesto que los protocolos de comunicaciones subyacentes no garantizan dicha solidez.

Formas de realización de la presente invención son resistentes contra dichos problemas haciendo idempotentes todos los mensajes e incidencias operativas que se intercambian entre el cliente del reconocimiento de voz 140 y el reconocedor en el lado del servidor 120. Una incidencia operativa es idempotente si múltiples ocurrencias de la misma incidencia tienen el mismo efecto que si produjera una ocurrencia única. Por lo tanto, si el cliente del reconocimiento de voz 140 detecta un fallo, dicho fallo en la transmisión de una orden al reconocedor en el lado del servidor 120, el cliente del reconocimiento de voz 140 puede retransmitir la orden, bien sea de inmediato, bien sea después de un periodo de espera. El cliente del reconocimiento de voz 140 y el reconocedor 120 pueden usar una interfaz de programa de aplicación (API) de mensajería que garantice que el reintento dejará al sistema 100 en un estado coherente.

En particular, la interfaz API para el flujo de voz 110 fuerza al cliente del reconocimiento de voz 140 a transmitir el flujo de voz 110 en segmentos. Cada segmento puede tener un ID único 304e en adición al índice de byte de inicio 304b (inicialmente 0 para el primer segmento) y un índice de byte final 304c o un tamaño de segmento. El reconocedor en el lado del servidor 120 puede confirmar que ha recibido un segmento transmitiendo, de nuevo, el índice de byte final del segmento, que normalmente debe ser igual al byte de inicio más el tamaño del segmento El índice de byte final transmitido por el servidor puede, sin embargo, ser un valor inferior si el servidor no pudiera efectuar la lectura del segmento de audio completo.

El cliente del reconocimiento de voz 140 transfiere, entonces, el siguiente segmento que se inicia en donde el reconocedor en el lado del servidor 120 lo dejó, por lo que el nuevo índice de byte de inicio es igual al índice de byte final reenviado por el reconocedor 120. Este proceso se repite para el flujo de voz 110 completo. Si se pierde un mensaje (en la ruta hacia o desde el servidor 118), el cliente del reconocimiento de voz 140 repite la transferencia. Si el reconocedor en el lado del servidor 120 no recibió anteriormente ese segmento de voz, en tal caso, el reconocedor en el lado del servidor 120 simplemente procesará los nuevos datos. Si, por el contrario, el reconocedor 120 procesó anteriormente ese segmento (tal como puede ocurrir si se perdieron los resultados en el retorno al cliente 106), entonces, el reconocedor 120 puede, a modo de ejemplo, acusar recibo del segmento y eliminarlo, de nuevo, sin procesamiento alguno.

Para el flujo de control 112, todos los mensajes de control 402a-c pueden reenviarse al servidor 118, puesto que cada uno de los mensajes puede contener un ID para la sesión operativa actual. En el caso del método de DecodeNext, el cliente del reconocimiento de voz 140 puede transmitir, como parte del método de DecodeNext, un identificador único de ejecución para identificar la llamada del método actual. El servidor 118 mantiene un registro de esos identificadores para determinar si el mensaje actual, que se está recibiendo en el flujo de control 112, es nuevo o si fue ya recibido y procesado. Si el mensaje actual es nuevo, en tal caso, el reconocedor 120 procesa el mensaje en la forma normal, según se describió con anterioridad. Si el mensaje actual fue procesado con anterioridad, entonces, el reconocedor 120 puede volver a entregar los resultados anteriormente reenviados, en lugar de volverlos a generar de nuevo.

Si uno de los mensajes de control 402a-c se envía al servidor 118 y el servidor 118 no acusa recibo del mensaje de control, el cliente 140 puede memorizar el mensaje de control. Cuando el cliente 140 tenga un segundo mensaje de control a enviar al servidor 118, el cliente 140 puede enviar, a la vez, el primer (sin confirmar) mensaje de control y el segundo mensaje de control al servidor 118. El cliente 140 puede, como alternativa, conseguir el mismo resultado combinando los cambios de estado representados por los primero y segundo mensajes de control en un mensaje de control único, que el cliente 140 puede transmitir luego al servidor 140. El cliente 140 puede combinar cualquier número de mensajes de control juntos en un mensaje de control único, de esta forma, hasta que dichos mensajes sean confirmados por el servidor 118. De modo similar, el servidor 118 puede combinar resultados del reconocimiento de voz, que no se hayan confirmado por el cliente 140, en resultados individuales en el flujo de resultados 122 hasta que dichos resultados sean confirmados por el cliente.

Entre las ventajas de la invención están una o más de las siguientes. Formas de realización de la presente invención permiten que el reconocimiento de voz se distribuya en cualquier lugar a través de Internet, sin requerir cualquier red especial. En particular, las técnicas aquí dadas a conocer pueden ponerse en práctica por intermedio de un protocolo de comunicación unidireccional, tal como HTTP, lo que permite la operación incluso en entornos restrictivos, en donde los clientes están limitados a participar solamente en comunicaciones salientes (unidireccionales). En consecuencia, formas de realización de la presente invención son de gran utilidad en conjunción con una amplia diversidad de redes sin ir en menoscabo de la seguridad. Además, las técnicas aquí dadas a conocer pueden reutilizar los mecanismos de seguridad de la web que ya existen (tales como SSL y, por extensión, HTTPS) para proporcionar unas comunicaciones seguras entre el cliente 106 y el servidor 118.

Según se indicó con anterioridad, una restricción común impuesta sobre los clientes es que sólo pueden usar un margen limitado de puertos salientes para comunicarse con los servidores externos. Formas de realización de la presente invención pueden ponerse en práctica en dichos sistemas multiplexando el flujo de voz 110 y el flujo de control 112 en un flujo único 114 que puede transmitirse a través de un puerto único.

Además, se puede requerir que sea encriptada la comunicación saliente. A modo de ejemplo, a los clientes se les suele permitir usar solamente el puerto de HTTPS encriptado seguro estándar (puerto 443). Formas de realización de la presente invención pueden ponerse en práctica por intermedio de un puerto HTTP estándar (no seguro) o un puerto HTTPS seguro para todas sus necesidades de comunicación, tanto para transferencia de audio 110 y como para flujo de control 112. En consecuencia, las técnicas aquí dadas a conocer pueden utilizarse en conjunción con sistemas que permitan a los clientes comunicarse utilizando sistemas HTTP no seguros y sistemas que requieren o permiten a los clientes comunicarse utilizando HTTPS seguros.

Las técnicas aquí dadas a conocer son también resistentes a fallos intermitentes de la red porque emplean un protocolo de comunicaciones en el que los mensajes son idempotentes. Esto es de especial utilidad cuando formas de realización de la presente invención se usan en conjunción con redes, tales como WANs, en cuyas redes son frecuentes las caídas de tensión y las sobretensiones. Aunque dichas incidencias operativas pueden causar fallos en los sistemas convencionales de reconocimiento de voz, en el lado del servidor, no afectan a los resultados obtenidos por las formas de realización de la presente invención (exceptuado posiblemente aumentando el tiempo de ida y retorno).

Formas de realización de la presente invención permiten que la voz 104 sea transmitida desde el cliente 106 al servidor 118 lo más rápido que permitiere la red 116, aun cuando el servidor 118 no pueda procesar continuamente esa voz. Además, el reconocedor 120, en el lado del servidor, puede procesar la voz desde la cola de espera de procesamiento 124 tan rápidamente como sea posible aun cuando la red 116 no pueda transmitir los resultados y/o la aplicación 108 no esté preparada para recibir los resultados. Estas y otras características de de formas de realización de la presente invención permiten que la voz y los resultados del reconocimiento de voz se transmitan y procesen, con la rapidez que permitieren los componentes individuales del sistema 100, de modo que los problemas con componentes individuales del sistema 100 tengan un impacto mínimo sobre el comportamiento operativo de los demás componentes del sistema 100.

Además, las formas de realización de la presente invención permiten al reconocedor 120, en el lado del servidor, procesar la voz tan rápidamente como sea posible, pero sin ir demasiado por delante de la aplicación del cliente 108. Según se describió anteriormente, la aplicación 108 puede utilizar los mensajes de control en el flujo de control 112

para emitir nuevas órdenes de configuración al reconocedor 120 que hacen que el reconocedor 120 se auto-reconfigure para reconocer la voz en el estado de configuración adecuado y para interrumpir temporalmente el reconocimiento, al producirse unas condiciones predeterminadas, de modo que la aplicación 108 pueda reconfigurar el estado del reconocedor 120 en una forma apropiada. Dichas técnicas permiten que se realice el reconocimiento de voz, tan rápidamente como sea posible, sin que se ponga en práctica utilizando el estado de configuración erróneo.

Ha de entenderse que aunque la invención se ha descrito anteriormente en términos de formas de realización particulares, las formas de realización anteriores son a título ilustrativo solamente y no limitan ni definen el alcance de la invención. Otras varias formas de realización incluyendo, sin limitación, la siguiente están también dentro del alcance de las reivindicaciones. A modo de ejemplo, los elementos y componentes aquí descritos pueden dividirse, además, en componentes adicionales o agruparse para formar menos componentes para realizar las mismas funcione.

Según se describió con anterioridad, varios métodos según las formas de realización de la presente invención pueden ponerse en práctica en paralelo, en su totalidad o en parte. Los expertos en esta materia apreciarán cómo realizar partes particulares de los métodos, aquí dados a conocer, para conseguir las ventajas indicadas, en varias combinaciones.

Las técnicas antes descritas pueden realizarse, a modo de ejemplo, en hardware, software, firmware o cualquiera de sus combinaciones. Dichas técnicas pueden poner en práctica en uno o más programas informáticos que se ejecutan en un ordenador programable incluyendo un procesador, un medio de almacenamiento legible por el procesador (incluyendo, a modo de ejemplo, memoria volátil y no volátil y/o elementos de almacenamiento), al menos un dispositivo de entrada y al menos un dispositivo de salida. Un código de programa puede aplicarse a la entrada introducida utilizando el dispositivo de entrada para realizar las funciones descritas y para generar una salida. La salida puede proporcionarse para uno o más dispositivos de salida.

Cada programa informático dentro del alcance de las reivindicaciones siguientes puede efectuarse en cualquier lenguaje de programación, tal como un lenguaje ensamblador, lenguaje de máquina, un lenguaje de programación de alto nivel o un lenguaje de programación orientado al objeto. El lenguaje de programación puede ser, a modo de ejemplo, un lenguaje de programación compilado o interpretado.

Cada uno de dichos programas informáticos puede ponerse en práctica en un producto de programa informático tangiblemente materializado en un dispositivo de almacenamiento legible por máquina para su ejecución por un procesador de ordenador. Las etapas del método de la invención pueden realizarse por un procesador de ordenador que ejecuta un programa tangiblemente materializado en un medio legible por ordenador para realizar funciones de la invención efectuando una entrada y generando la salida. Procesadores adecuados incluyen, a modo de ejemplo, microprocesadores de uso general y para fines especiales. En términos generales, el procesador recibe instrucciones y datos desde una memoria de solamente lectura y/o una memoria de acceso aleatorio. Los dispositivos de memorización adecuados para materializar tangiblemente las instrucciones de programas informáticos incluyen, a modo de ejemplo, todas las formas de memoria no volátil, tales como dispositivos de memoria de semiconductores, incluyendo dispositivos de memoria EPROM, EEPROM y de memoria instantánea; discos magnéticos tales como discos duros internos y discos extraíbles; discos magneto-ópticos y CD-ROMs. Cualquiera de los dispositivos anteriores puede complementarse por, o incorporarse en, circuitos ASICs de diseño especial (circuitos integrados específicos de la aplicación) o FPGAs (Conjuntos de puertas programables en situ).Un ordenador puede, en general, recibir también programas y datos desde un medio de memorización tal como un disco interno (no ilustrado) o un disco extraíble. Estos elementos se encontrarán también en un ordenador de sobremesa convencional o en un ordenador de estación de trabajo así como en otros ordenadores adecuados para ejecutar programas informáticos que ponen en práctica los métodos aquí descritos, que pueden utilizarse en conjunción con cualquier motor de impresión digital o motor de marcado, monitor de presentación visual u otro dispositivo de salida de trama capaz de producir elementos de imagen, *pixels*, de color o de escala de grises en papel, película, pantalla de presentación visual u otro medio de salida.

REIVINDICACIONES

1. Un método puesto en práctica por ordenador que comprende:

(A) al nivel de un cliente (106), la transmisión de un flujo de voz y de un flujo de control a un servidor de reconocimiento de voz (118) utilizando un Protocolo de Transferencia de Hipertexto, HTTP, que tiene un primer periodo de temporización;

(B) al nivel del servidor de reconocimiento de voz, la utilización de un motor de reconocimiento de voz automático para iniciar el reconocimiento del flujo de voz;

(C) al nivel del cliente, la transmisión de una primera demanda de un resultado de reconocimiento de voz al servidor de reconocimiento de voz utilizando HTTP y

(D) al nivel de servidor de reconocimiento de voz, la transmisión de una notificación al cliente indicando que ningún resultado de reconocimiento de voz se hizo disponible dentro de un segundo periodo de temporización que difiere del primer periodo de temporización y

(E) al nivel del cliente, en respuesta a la recepción de la notificación, la transmisión de una segunda demanda del resultado de reconocimiento de voz al servidor de reconocimiento de voz utilizando HTTP.

2. El método según la reivindicación 1, que comprende, además:

(F) al nivel del servidor de reconocimiento de voz (118), el reconocimiento de una primera parte del flujo de voz para producir un primer resultado de reconocimiento de voz y

(G) la transmisión del primer resultado de reconocimiento de voz al cliente (106) utilizando HTTP en respuesta a la segunda demanda.

3. El método según la reivindicación 2, en donde (G) comprende:

(G) (1) la determinación de si cualesquiera resultados de reconocimiento de voz están disponibles;

(G) (2) si no están disponibles resultados del reconocimiento de voz, volver a (G)(1);

(G) (3) si no es así, la transmisión del primer resultado de reconocimiento de voz al cliente (106).

4. El método según la reivindicación 3, en donde el servidor de reconocimiento de voz (118) realiza (F) y (G) en paralelo.

5. El método según la reivindicación 1, en donde (A) comprende la transmisión del flujo de voz y del flujo de control utilizando un Protocolo de Transferencia de Hipertexto sobre la Capa de Conexión Segura HTTPS, y en donde (C) comprende la transmisión de la primera demanda utilizando HTTPS.

6. Un sistema que comprende un dispositivo cliente (106) y un servidor de reconocimiento de voz (118):

en donde el dispositivo cliente comprende:

medios (110, 112) para efectuar la transmisión de un flujo de voz y de un flujo de control al servidor de reconocimiento de voz utilizando un Protocolo de Transferencia de Hipertexto, HTTP, que tiene un primer periodo de temporización;

medios para efectuar la transmisión de una primera demanda de un resultado de reconocimiento de voz al servidor de reconocimiento de voz utilizando HTTP y

en donde el servidor de reconocimiento de voz comprende:

medios para, utilizando un motor de reconocimiento de voz automático (120, 218), iniciar el reconocimiento del flujo de voz;

medios para efectuar la transmisión de una notificación al dispositivo cliente indicando que ningún resultado de reconocimiento de voz se hizo disponible dentro de un segundo periodo de temporización que difiere del primer periodo de temporización y

en donde el dispositivo cliente comprende, además, medios para efectuar, en respuesta a la recepción de la notificación, la transmisión de una segunda demanda del resultado del reconocimiento de voz al servidor de

reconocimiento de voz utilizando HTTP.

7. Un método puesto en práctica por ordenador realizado por un servidor de reconocimiento de voz (118), cuyo método comprende:

(A) la recepción de un flujo de voz y de un flujo de control desde un cliente (106) utilizando un Protocolo de Transferencia de Hipertexto, HTTP, que tiene un primer periodo de temporización;

(B) la utilización de un motor de reconocimiento de voz automático para iniciar un reconocimiento del flujo de voz;

(C) la recepción de una primera demanda de un resultado de reconocimiento de voz desde el cliente utilizando HTTP y

(D) la transmisión de una notificación al cliente indicando que ningún resultado de reconocimiento de voz se hizo disponible dentro de un segundo periodo de temporización que difiere del primer periodo de temporización.

8. El método según la reivindicación 7 que comprende, además:

(E) la recepción de una segunda demanda del resultado de reconocimiento de voz desde el cliente (106) utilizando HTTP;

(F) el reconocimiento de una primera parte del flujo de voz para producir un primer resultado de reconocimiento de voz y

(G) la transmisión del primer resultado de reconocimiento de voz al cliente utilizando HTTP en respuesta a la segunda demanda.

9. Un aparato que comprende:

medios para efectuar la recepción de un flujo de voz (110) y de un flujo de control (112) desde un cliente (106) utilizando un Protocolo de Transferencia de Hipertexto, HTTP, que tiene un primer periodo de temporización;

medios para, utilizando un motor de reconocimiento de voz automático, iniciar un reconocimiento del flujo de voz;

medios para efectuar la recepción de una primera demanda de un resultado de reconocimiento de voz desde el cliente utilizando HTTP; y

medios para efectuar la transmisión de una notificación al cliente indicando que ningún resultado de reconocimiento de voz se hizo disponible dentro de un segundo periodo de temporización que difiere del primer periodo de temporización.

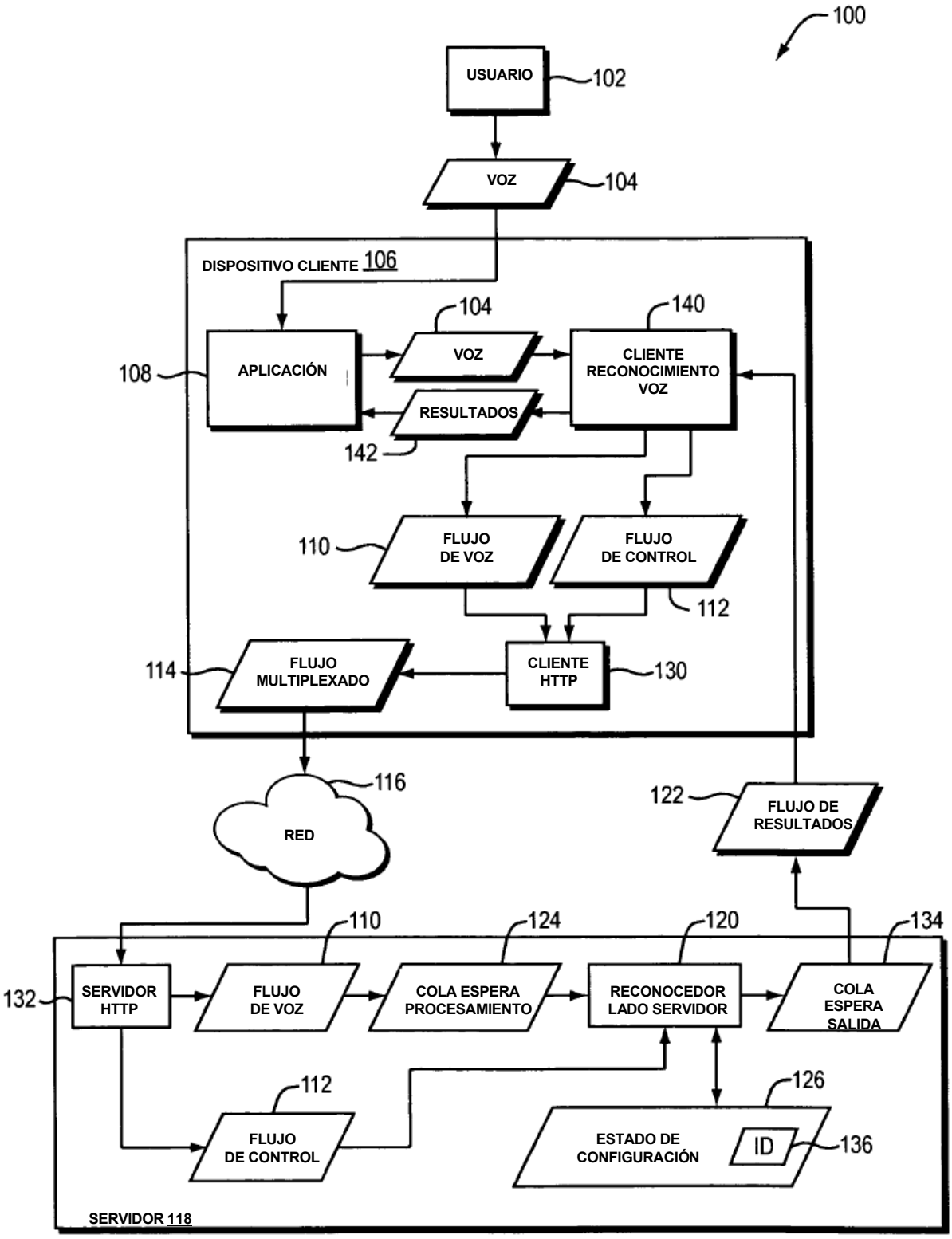
100
USUARIO
102
VOZ
104
DISPOSITIVO CLIENTE 106
104
140
108
APLICACIÓN
VOZ
CLIENTE RECONOCIMIENTO VOZ
RESULTADOS
142
110
FLUJO DE VOZ
FLUJO DE CONTROL
112
114
FLUJO MULTIPLEXADO
CLIENTE HTTP
130
116
RED
122
FLUJO DE RESULTADOS
110
124
120
134
132
SERVIDOR HTTP
FLUJO DE VOZ
COLA ESPERA PROCESAMIENTO
RECONOCEDOR LADO SERVIDOR
COLA ESPERA SALIDA
112
126
FLUJO DE CONTROL
ESTADO DE CONFIGURACIÓN
ID
136
SERVIDOR 118

FIG. 1

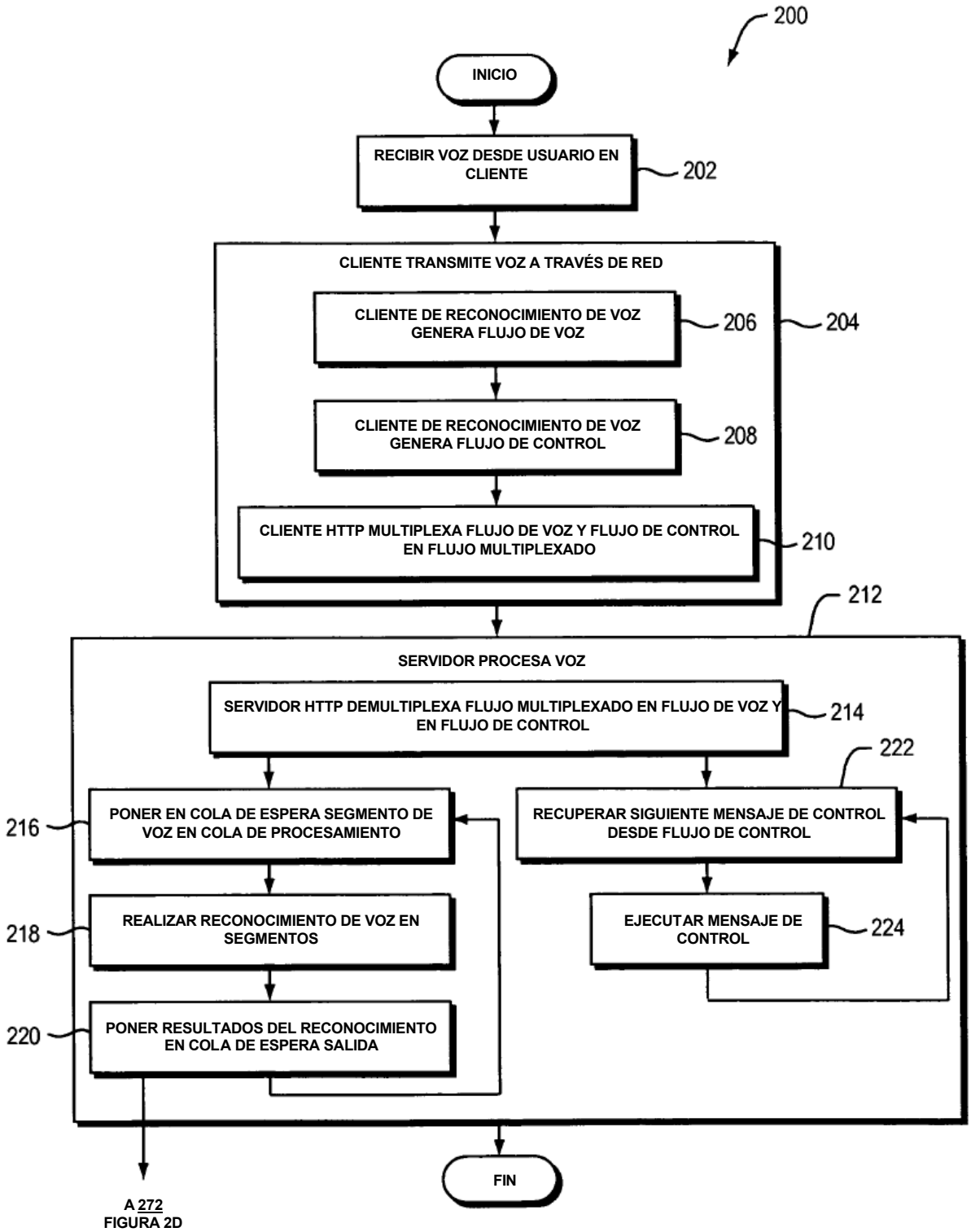
200
INICIO
RECIBIR VOZ DESDE USUARIO EN CLIENTE
202
CLIENTE TRANSMITE VOZ A TRAVÉS DE RED
CLIENTE DE RECONOCIMIENTO DE VOZ GENERA FLUJO DE VOZ
206
204
CLIENTE DE RECONOCIMIENTO DE VOZ GENERA FLUJO DE CONTROL
208
CLIENTE HTTP MULTIPLEXA FLUJO DE VOZ Y FLUJO DE CONTROL EN FLUJO MULTIPLEXADO
210
212
SERVIDOR PROCESA VOZ
SERVIDOR HTTP DEMULTIPLEXA FLUJO MULTIPLEXADO EN FLUJO DE VOZ Y EN FLUJO DE CONTROL
214
222
216
PONER EN COLA DE ESPERA SEGMENTO DE VOZ EN COLA DE PROCESAMIENTO
RECUPERAR SIGUIENTE MENSAJE DE CONTROL DESDE FLUJO DE CONTROL
218
REALIZAR RECONOCIMIENTO DE VOZ EN SEGMENTOS
EJECUTAR MENSAJE DE CONTROL
224
220
PONER RESULTADOS DEL RECONOCIMIENTO EN COLA DE ESPERA SALIDA
FIN
A 272
FIGURA 2D

FIG. 2A

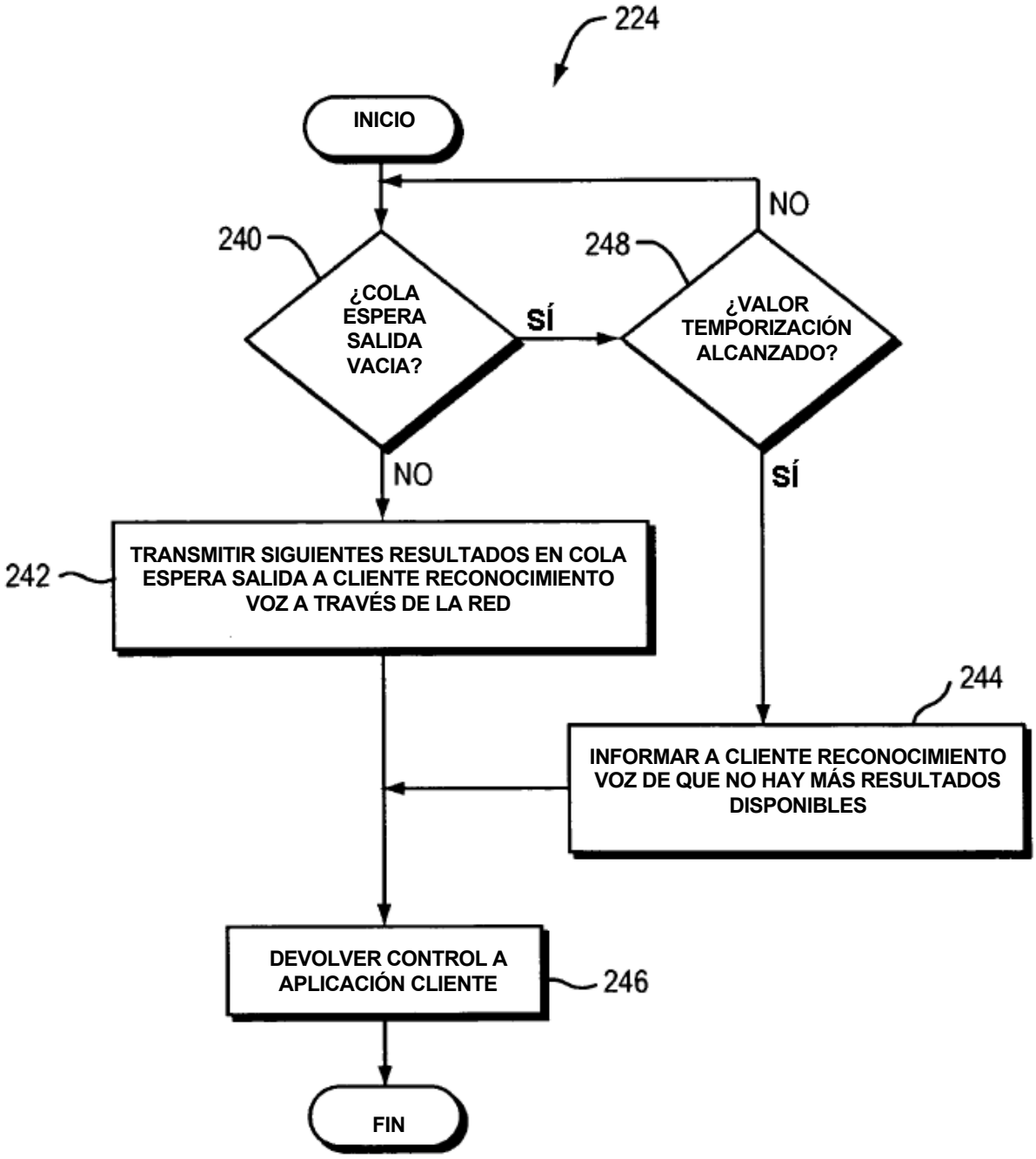
224
INICIO
NO
240
248
¿COLA ESPERA SALIDA VACIA?
SÍ
¿VALOR TEMPORIZACIÓN ALCANZADO?
NO
SÍ
242
TRANSMITIR SIGUIENTES RESULTADOS EN COLA ESPERA SALIDA A CLIENTE RECONOCIMIENTO VOZ A TRAVÉS DE LA RED
244
INFORMAR A CLIENTE RECONOCIMIENTO VOZ DE QUE NO HAY MÁS RESULTADOS DISPONIBLES
DEVOLVER CONTROL A APLICACIÓN CLIENTE
246
FIN

FIG. 2B

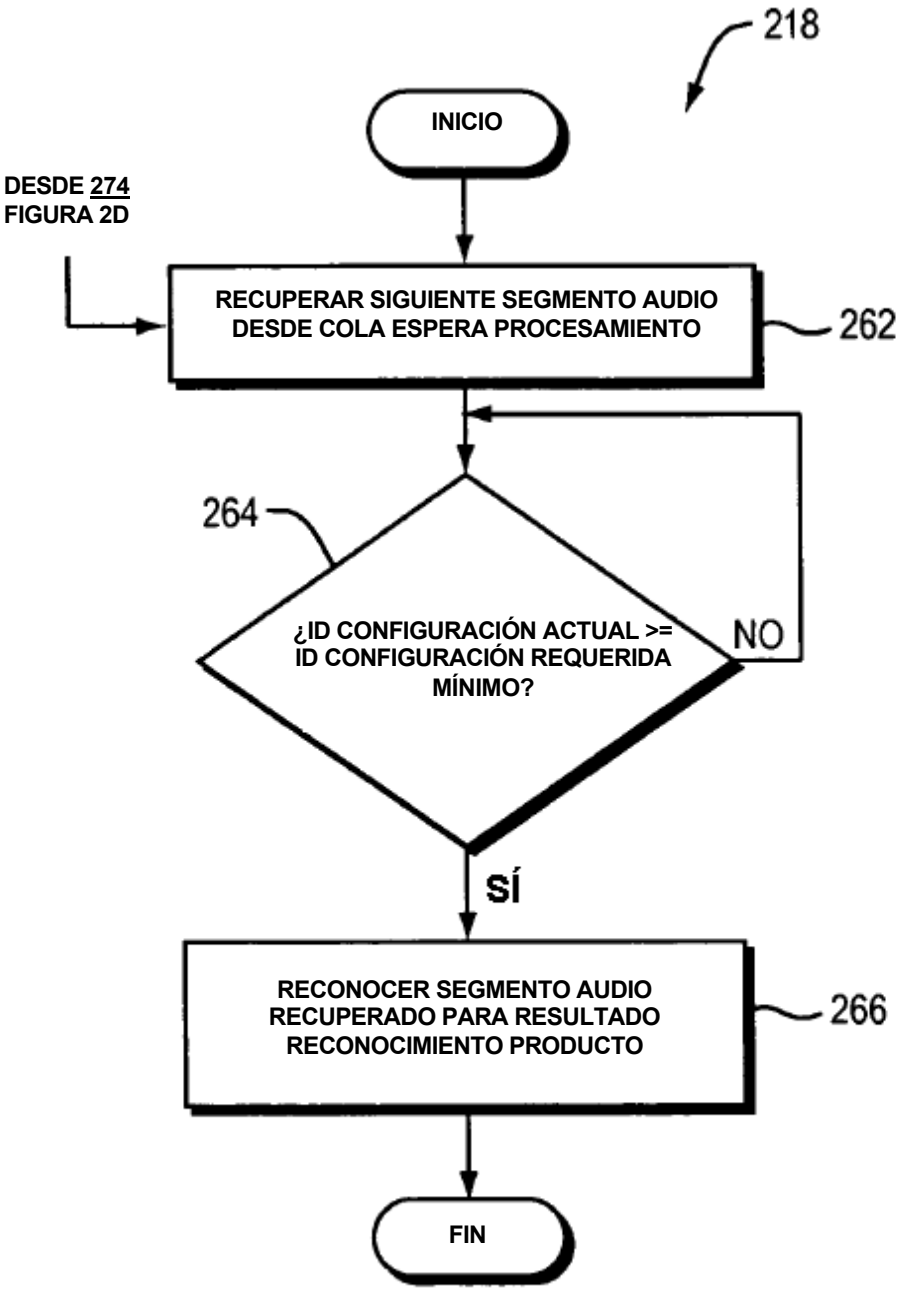
218
INICIO
DESDE 274
FIGURA 2D
RECUPERAR SIGUIENTE SEGMENTO AUDIO DESDE COLA ESPERA PROCESAMIENTO
262
264
¿ID CONFIGURACIÓN ACTUAL >= ID CONFIGURACIÓN REQUERIDA MÍNIMO?
NO
SÍ
RECONOCER SEGMENTO AUDIO RECUPERADO PARA RESULTADO RECONOCIMIENTO PRODUCTO
266
FIN

FIG. 2C

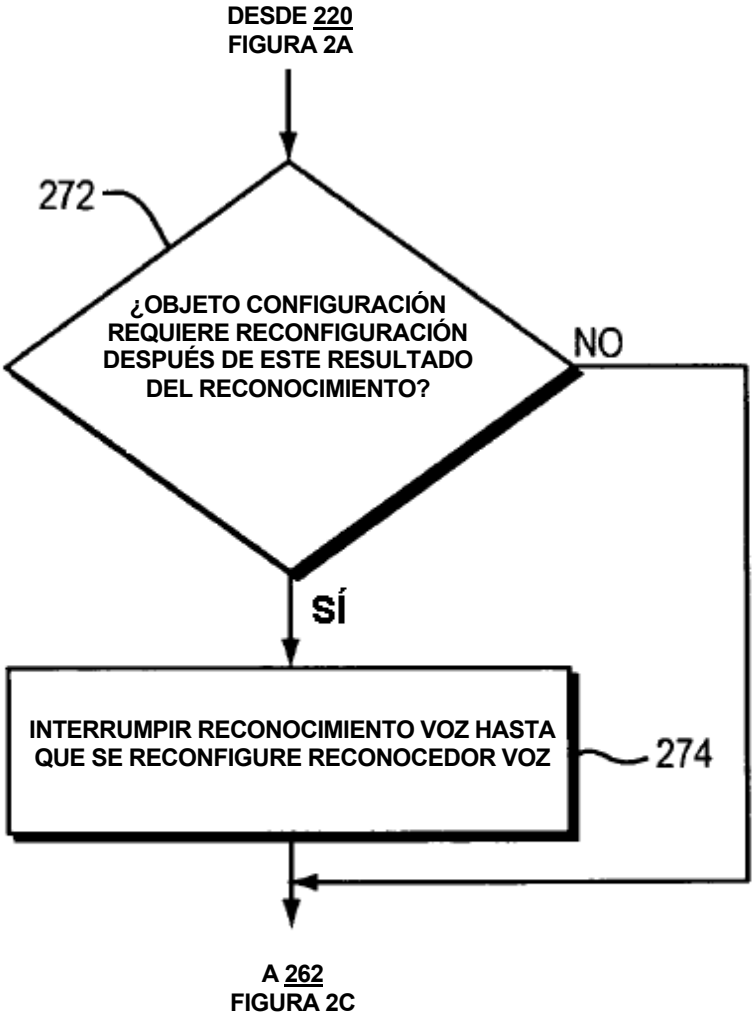
DESDE 220
FIGURA 2A
272
¿OBJETO CONFIGURACIÓN REQUIERE RECONFIGURACIÓN DESPUÉS DE ESTE RESULTADO DEL RECONOCIMIENTO?
NO
SÍ
INTERRUMPIR RECONOCIMIENTO VOZ HASTA QUE SE RECONFIGURE RECONOCEDOR VOZ
274
A 262
FIGURA 2C

FIG. 2D

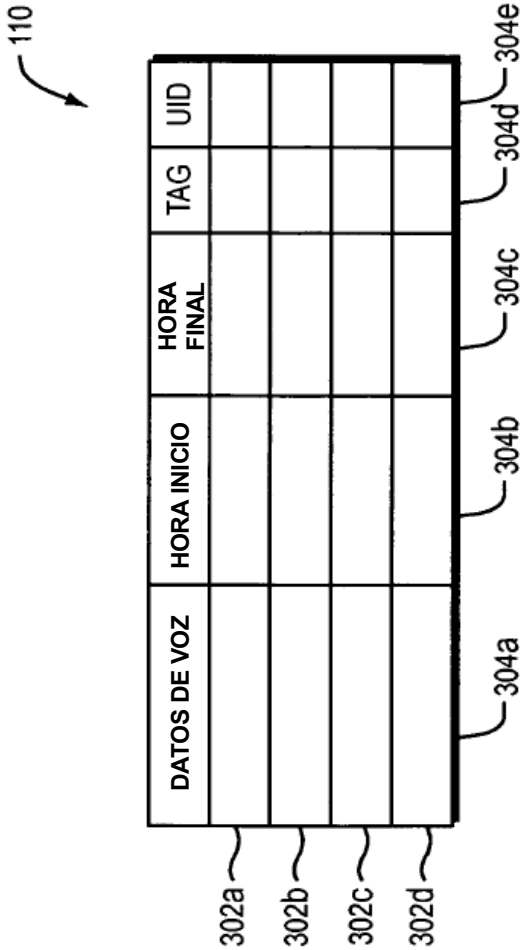
110
DATOS DE VOZ
HORA INICIO
HORA FINAL
TAG
UID
302a
302b
302c
302d
304a
304b
304c
304d
304e

FIG. 3

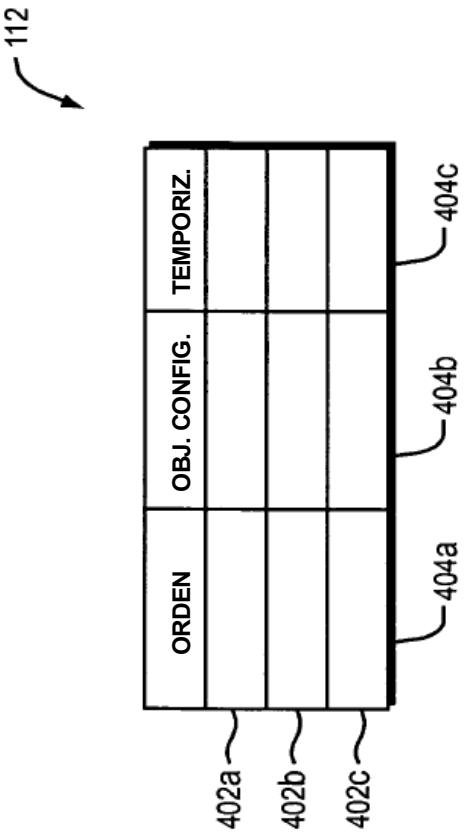
112
ORDEN
OBJ. CONFIG.
TEMPORIZ.
402a
402b
402c
404a
404b
404c

FIG. 4