(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 386 673**

(51) Int. CI.:
G10L 15/26 (2006.01)
G11B 27/031 (2006.01)
H04M 3/42 (2006.01)
G10L 15/28 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **08774723 .4**
(96) Fecha de presentación: **03.07.2008**
(97) Número de publicación de la solicitud: **2311031**
(97) Fecha de publicación de la solicitud: **20.04.2011**

(54) Título: **Procedimiento y dispositivo de conversión de voz**

(45) Fecha de publicación de la mención BOPI:
24.08.2012

(45) Fecha de la publicación del folleto de la patente:
24.08.2012

(73) Titular/es:
Mobiter Dicta Oy
Topeliuksenkaut 3a a5
00260 Helsinki, FI

(72) Inventor/es:
KURKI-SUONIO, Risto y
COTTON, Andrew

(74) Agente/Representante:
Carpintero López, Mario

DESCRIPCIÓN

Procedimiento y dispositivo de conversión de voz

Campo de la invención

La presente invención se refiere, en general, a dispositivos electrónicos y redes de comunicaciones. En particular, aunque no exclusivamente, la invención se refiere a aplicaciones de conversión de voz en texto.

Antecedentes de la invención

La tendencia actual en terminales portátiles, por ejemplo, manuales, impulsa la evolución fuertemente hacia interfaces de usuario intuitivas y naturales. Además de texto, imágenes y sonido (por ejemplo, la voz) se pueden grabar en un terminal, ya sea para transmisión o para controlar una funcionalidad preferida local o remota (es decir, basada en la red). Además, la información de carga útil puede ser transferida a través de las redes celulares y fijas adyacentes tales como Internet, como datos binarios que representan el texto, sonido, imágenes y vídeo subyacente. Aparatos de módem en miniatura, tales como terminales móviles o PDAs (Asistentes Digitales Personales) pueden así llevar medios de entrada de control versátiles tales como una pantalla/teclado, un micrófono, diferentes sensores de movimiento o presión, etc. para proporcionar a los usuarios de los mismos una UI (interfaz de usuario) realmente capaz de soportar los mecanismos de almacenamiento y comunicación de datos muy diversificados.

A pesar del salto en la tecnología de comunicación e información en curso, también algunas soluciones de almacenamiento de datos más tradicionales, tales como aparatos de dictado parecen mantener un valor de uso considerable, especialmente en campos especializados como el derecho y las ciencias médicas, en las que los documentos se crean regularmente sobre la base de las discusiones verbales y reuniones, por ejemplo. Es probable que la comunicación verbal siga siendo el procedimiento más rápido y conveniente de expresión para la mayoría de la gente, y dictando una nota en lugar de escribirla puede lograrse un ahorro de tiempo considerable. Este tema también tiene un aspecto de dependencia del idioma; la escritura china o japonesa, obviamente, requieren más tiempo para escribir que la mayoría de los idiomas occidentales, por ejemplo. Además, las máquinas de dictar y sus homólogas modernas, tales como sofisticados terminales móviles y PDAs con opción de grabación de sonido pueden utilizarse hábilmente junto con otras tareas, por ejemplo, mientras se tiene una reunión o se conduce un coche, mientras que la escritura manual requiere normalmente una parte importante de la atención de la persona que la realiza y definitivamente no se puede realizar si se conduce un coche, etc.

Hasta los últimos pocos años, sin embargo, los aparatos de dictado no han servido para todas las necesidades del público tan bien; es cierto que la información puede ser fácilmente almacenada, incluso en tiempo real con sólo grabar la señal de voz a través de un micrófono, pero a menudo la forma de archivo final es textual y alguien, por ejemplo, una secretaria, ha recibido la orden de limpiar manualmente y convertir la señal grabada de sonido en bruto en un registro final en un medio diferente. Esta disposición, lamentablemente, requiere una gran cantidad de trabajo adicional de conversión que consume tiempo. Otro problema importante asociado con las máquinas de dictado surge de su origen analógico e interfaz de usuario simple; la modificación de la voz que ya está almacenada es muy complicada y con muchos dispositivos, todavía utilizando cinta magnética como medio de almacenamiento, ciertas operaciones de edición, tal como la inserción de una porción de voz completamente nueva dentro de la señal original almacenada no pueden realizarse. Mientras tanto, las máquinas modernas de dictado que utilizan chips/tarjetas de memoria pueden comprender opciones de edición de la voz limitadas, pero la posible utilización todavía está disponible sólo a través de una interfaz de usuario bastante incómoda que comprende sólo una pantalla LDC (Pantalla de Cristal Líquido) de mínimo tamaño y calidad, etc. La transferencia de datos de voz almacenados a otro dispositivo, requiere a menudo manipulación, es decir, el medio de almacenamiento (cinta/tarjeta de memoria) debe moverse físicamente.

Sistemas computarizados de reconocimiento de la voz han estado disponibles para una persona experta en la materia desde hace algún tiempo. Estos sistemas están normalmente implementados como características internas de aplicaciones específicas (incorporadas en un procesador de texto, por ejemplo, Microsoft Word versión XP), aplicaciones en solitario, o plugins de aplicación para un ordenador de escritorio normal. El proceso de reconocimiento de la voz implica una serie de etapas que están básicamente presentes en todos los algoritmos existentes, véase la figura 1 para ilustración de un ejemplo particular. A saber, la señal de origen de la voz emitida por una persona que habla primero es capturada 102 a través de un micrófono o un transductor correspondiente y es convertida en forma digital con un procesamiento previo necesario 104 que puede referirse a procesamiento dinámico, por ejemplo. A continuación, la señal digitalizada se entra a un motor de reconocimiento de la voz 106 que divide la señal en elementos más pequeños como fonemas basados en la

extracción de características sofisticadas y procedimientos de análisis. El software de reconocimiento también puede estar adaptado a cada usuario 108, es decir, la configuración del software es específica para el usuario. Finalmente, los elementos reconocidos que forman la salida del motor de reconocimiento de voz, por ejemplo, información y/o texto de control, se utilizan como entrada 110 para otros fines; simplemente se puede mostrar en la pantalla, almacenar en una base de datos, traducir a otro idioma, usar para ejecutar una funcionalidad predeterminada, etc., tal como se describe en la publicación EP0664636, "Sistema de conferencia de audio", R.A. Sharman et al., 26.07.1995.

La publicación US6266642 divulga una unidad portátil dispuesta para realizar la traducción del idioma hablado para facilitar la comunicación entre dos entidades que no tienen ningún idioma común. El propio dispositivo contiene todo el hardware y el software necesarios para ejecutar el proceso de traducción o simplemente actúa como una interfaz remota que, inicialmente, dirige, mediante la utilización de un teléfono o una videoconferencia, la voz de entrada en la unidad de traducción para su procesamiento, y más tarde recibe el resultado de la traducción para síntesis de la voz local. La solución también comprende una etapa de procesamiento en la que se minimizan los fallos de reconocimiento de la voz mediante la creación de una serie de reconocimientos candidatos o hipótesis de que el usuario puede, a través de una interfaz de usuario, seleccionar la opción correcta, o simplemente confirmar la selección predefinida.

A pesar de los muchos avances de las disposiciones antes mencionadas y otras sugerencias de disposiciones de la técnica anterior para superar las dificultades encontradas en el reconocimiento de la voz y/o los procesos de traducción automática, algunos problemas siguen sin resolverse, especialmente en relación con los dispositivos móviles. Los problemas asociados con las máquinas tradicionales de dictado ya fueron descritos anteriormente. Además, muchos grupos de usuarios especiales, tal como las personas con discapacidad, incluidos los usuarios ciegos, han sido muy comúnmente olvidados en el diseño de la interfaz de usuario de los dispositivos de reconocimiento de voz más sofisticados, conversión de voz a texto, o de traducción y servicios asociados a las interfaces de usuario, dependiendo generalmente en gran medida en la orientación de procesos y características de visualización de datos en una pantalla de bajo contraste/baja resolución de tamaño pequeño, por ejemplo.

Aún más, muchas aplicaciones capaces de registro y reconocimiento de voz se han adaptado para capturar y procesar de manera totalmente autónoma la señal de entrada de audio en un objetivo predeterminado después de recibir una petición de procesamiento inicial que puede referirse a una señal creada presionando un botón de iniciación correspondiente en la interfaz de usuario del dispositivo asociado, por ejemplo. Sin embargo, aunque varias funcionalidades totalmente automatizadas en general son bienvenidas, ya que pueden superar la necesidad de ajustes manuales más exhaustivos o de control continuo, las soluciones automatizadas no siempre proporcionan una precisión similar a las alternativas manuales o semiautomáticas, y, lo que es igualmente importante, las soluciones automatizadas a veces ejercen presión sobre el usuario de las mismas cuando el usuario se ve obligado a actuar anormalmente en una situación algo básica, es decir, la solución obliga al usuario a adaptarse a la situación de uso del dispositivo en particular aplicado, que puede diferir de la manera innata verdadera natural de hacer la tarea asociada, tales como el dictado. Esto puede resultar en una experiencia incómoda del usuario y molesta que finalmente conduce al usuario a abstenerse subliminalmente de la utilización del dispositivo para tal fin.

Sumario de la invención

El objetivo de la invención es aliviar al menos algunos de los defectos antes mencionados que se encuentran en las disposiciones actuales de archivo de voz y conversión de voz a texto.

El objetivo se consigue mediante una solución en la que un dispositivo electrónico, por ejemplo, un ordenador de sobremesa, portátil u ordenador de mano, un terminal móvil tal como un teléfono GSM/UMTS/CDMA, una PDA, o una máquina de dictado, opcionalmente equipada con un adaptador o transceptor de comunicaciones inalámbricas, se proporciona con una funcionalidad para obtener información de control desde el usuario del dispositivo durante una operación de captura de la señal de voz para cultivar el reconocimiento de voz en curso o posterior, en particular conversión de voz a texto, procedimiento que está al menos parcialmente automatizado.

Por consiguiente, en un aspecto de la invención, se proporciona un dispositivo electrónico como se expone en la reivindicación 1 para facilitar el procedimiento de conversión de voz a texto, que comprende:

- unos medios de entrada de voz para obtener una señal de voz digital,

- unos medios de entrada de control para comunicar un comando de control relativo a la señal de voz digital, mientras se obtiene la señal de voz digital,

- unos medios de procesamiento para asociar temporalmente el comando de control con un instante de tiempo que corresponde sustancialmente en la señal de voz digital a la que se dirige el comando de control,

en el que el comando de control determina uno o más signos de puntuación, símbolos u otros elementos de control que implican la manipulación de texto, para estar físicamente, tal como en el caso de dichos signos de puntuación y símbolos, o al menos lógicamente, a través de la manipulación de texto en el caso de dichos otros elementos de control, situados en una ubicación en el texto correspondiente al instante de comunicación relativo a la señal de voz digital, para procurar la voz con el procedimiento de conversión de texto localmente, en cuyo caso el dispositivo también comprende un motor de reconocimiento de voz para realizar tareas de conversión de voz a texto, o de forma remota, en cuyo caso el dispositivo electrónico también comprende unos medios de transferencia de datos para el envío de datos digitales que representan la señal de voz digital y el comando de control remoto a una entidad para la conversión, o mediante un procedimiento de conversión compartido entre el dispositivo electrónico y la entidad remota, en cuyo caso el dispositivo electrónico también comprende al menos una parte del motor de reconocimiento de voz y dichos medios de transferencia de datos.

El dispositivo así puede colocar los elementos en el resultado de la conversión (texto), tal como se indica mediante el momento de su adquisición en relación con la señal de voz, pero opcionalmente también inicia una o más acciones predeterminadas diferentes, o "tareas", tal como una pausa en la grabación de longitud predeterminada, en respuesta a la obtención del comando de control. Las acciones pueden iniciarse inmediatamente después de obtener el comando o de una manera retardada, por ejemplo, con un retardo predeterminado.

El dispositivo electrónico puede, además, en otro aspecto, comprender una ayuda especial destinada especialmente para las personas ciegas o con dificultades de visión, proporcionando funcionalidad para confirmar, según el criterio predeterminado, porciones de texto inciertas convertidas de voz a texto a través de una serie de opciones clasificadas entre sí.

Por lo tanto, el dispositivo electrónico para llevar a cabo al menos parte del procedimiento de conversión de voz a texto puede comprender adicionalmente

- unos medios de procesamiento o transferencia de datos para obtener al menos una conversión parcial de voz a texto que incluye una porción convertida, tales como una o más palabras o frases, que comprende múltiples, dos o más, opciones de resultado de conversión seleccionables por el usuario,

- unos medios de salida de audio para reproducir de manera audible una o más de dichas opciones para dicha porción, y

- unos medios de entrada de control para comunicar una selección del usuario de una de dichas múltiples opciones seleccionables por el usuario para permitir la confirmación de un resultado de la conversión deseada para dicha porción.

En ambos aspectos, el dispositivo puede actuar como un terminal remoto de reconocimiento de voz/motor de conversión de voz a texto que reside en una conexión de comunicaciones. Alternativamente, el dispositivo puede incluir en sí mismo el motor sin necesidad de elementos externos de contacto. También una solución mixta con reparto de tareas es posible, tal como se describirá más adelante.

El usuario puede, sobre la base de la reproducción audible, que no impide el uso de otro tipo de medios de reproducción adicionales o alternativos, tales como medios visuales o táctiles, seleccionar una conversión adecuada resultado de opciones múltiples. Las opciones pueden ser clasificadas y se reproducen de acuerdo a su relevancia preliminar, por ejemplo. Como consecuencia, si el usuario escucha la opción correcta en primer lugar, que preferiblemente sucede muy a menudo, inmediatamente puede confirmar la selección en lugar de escuchar también otras opciones, inevitablemente inferiores. Para situaciones en las que ninguna de las opciones es correcta, unos medios de interfaz de usuario predeterminados pueden ser seleccionados para ignorar todas las opciones representadas, mediante lo cual el dispositivo puede adaptarse para registrar la porción de voz relacionada una vez más para el reconocimiento repetido y, opcionalmente, la selección del usuario de un texto adecuado alternativo.

Preferiblemente, la(s) porción(es) mencionada(s) se selecciona(n) de manera que cubre(n) sólo una pequeña parte de todo el resultado de la conversión tal que el usuario no tiene que volver a verificar y verificar

manualmente el resultado de cada segundo de conversión único, que se puede garantizar proporcionando las únicas opciones para las porciones más fiables de las palabras o frases, por ejemplo. El número de tales porciones más fiables seleccionadas para la confirmación del usuario puede ser restringido en términos absolutos o por unidad de tiempo predeterminado y/o cantidad de texto, por ejemplo.

La reproducción puede utilizar un sintetizador de texto a voz que aplica un modelo de producción de voz, tal como un modelo de síntesis formante, y/o alguna otra solución, tal como un banco de muestras, es decir, la voz grabada. Las preferencias de reproducción pueden ser ajustables. Por ejemplo, la síntesis de voz, la velocidad o el volumen pueden ser seleccionables por el usuario, dependiendo de la realización.

En ambos aspectos, los medios de control de entrada, por ejemplo, pueden referirse a uno o más botones, llaves, mandos, una pantalla táctil, medios ópticos de entrada, controlador de reconocimiento de voz, etc. estando conectados al menos funcionalmente al dispositivo. Los medios de entrada de voz pueden referirse a uno o más micrófonos o conectores para micrófonos externos, y medios de conversión A/D, o a una interfaz para la obtención la señal ya de forma digital de voz desde una fuente externa, tal como un micrófono digital suministrado con un transmisor. Los medios de procesamiento pueden referirse a uno o más microprocesadores, microcontroladores, chips lógicos programables, procesadores de señales digitales, etc. Los medios de transferencia de datos pueden referirse a uno o más cables o interfaces de datos inalámbricos, tales como transceptores, a sistemas o dispositivos externos. Los medios de salida de audio pueden referirse a uno o más altavoces o conectores para altavoces externos u otros medios de salida de audio, por ejemplo.

El dispositivo electrónico comprende opcionalmente una interfaz de usuario que permite al usuario, mediante visualización o por otros medios, modificar la señal de voz antes de que se exponga al reconocimiento de voz real y procesos opcionales, por ejemplo, traducción. Además, en algunas realizaciones de la invención, la comunicación entre el dispositivo y una entidad externa, por ejemplo, un servidor de red que reside en la red a la que el dispositivo tiene acceso, puede desempeñar un papel importante. El dispositivo y la entidad externa pueden configurarse para dividir la ejecución de la voz para la conversión de texto y nuevas medidas sobre la base de una serie de valores de parámetros ventajosamente definibles por el usuario relativos a, entre otros, posibles factores locales/remotos de procesamiento/carga de memoria, estado de la batería, existencia de otras tareas prioritarias de los mismos, y ancho de banda de transmisión disponible, aspectos relacionados con el coste, tamaño/duración de la fuente de señal de voz, etc. El dispositivo y la entidad externa pueden incluso negociar una escenario de cooperación adecuado en tiempo real basado en sus condiciones actuales, es decir, el intercambio de tareas es un proceso dinámico. También estas cuestiones opcionales se discuten a continuación con más detalle. El proceso de conversión en su conjunto, por lo tanto, puede ser interactivo entre el usuario del dispositivo, el dispositivo en sí y la entidad externa. Además, el proceso de reconocimiento de voz se puede personalizar en relación con cada usuario, es decir, el motor de reconocimiento puede configurarse por separado o estar capacitado para adaptarse a sus características de voz.

En un escenario, el dispositivo electrónico puede ser un dispositivo móvil operable en una red de comunicaciones inalámbrica que comprende unos medios de entrada de voz para recibir la voz y convertir la voz en una señal representativa de voz digital, unos medios de entrada de control para comunicar un comando de edición relacionado con la señal de voz digital, unos medios de procesamiento para la realización de una tarea de edición de la señal de voz digital en respuesta al comando de edición recibido, al menos parte de un motor de reconocimiento de voz para la realización de tareas de una señal de voz digital para la conversión de texto, y un transceptor para el intercambio de información relativa al la señal digital de voz y la conversión de voz a texto del mismo con una entidad externa funcionalmente conectada a dicha red de comunicaciones inalámbricas.

En el escenario anterior, el comando de edición y la tarea asociada pueden estar relacionados, pero no se limitan, a una de las siguientes opciones: eliminación de una porción de la señal de voz, inserción de una nueva porción de voz en la señal de voz, sustitución de una porción en el señal de voz, cambio en la amplitud de la señal de voz, cambio en el contenido espectral de la señal de voz, volver a grabar una porción de la señal de voz. Preferiblemente, el dispositivo móvil incluye medios de visualización para visualizar la señal de voz digital, de manera que los comandos de edición pueden referirse a la(s) porción(es) de señal visualizada(s).

El motor de reconocimiento de voz puede comprender un marco, lógica de análisis, por ejemplo, en una forma de hardware y/o software a medida que se requiere para la ejecución de al menos parte del proceso global de la conversión de voz a texto a partir de voz en forma digital. Un proceso de reconocimiento de voz se refiere generalmente a un análisis de una señal de audio (que comprende la voz) sobre la base del cual puede la señal puede dividirse en porciones más pequeñas y clasificar las porciones. El reconocimiento de voz permite así y forma (al menos) una parte importante del procedimiento completo de conversión de voz a texto de la invención, aunque la salida simple del motor de reconocimiento de voz también podría ser algo más que el texto que representa textualmente la voz hablada; por ejemplo, en aplicaciones de control de voz, el motor de

reconocimiento de voz asocia la voz de entrada con una serie de comandos predeterminados que el dispositivo del servidor está configurado para ejecutar. El proceso de conversión incluye típicamente una pluralidad de etapas y, por lo tanto, el motor puede llevar a cabo sólo una parte de las etapas o, alternativamente, la señal de voz se puede dividir en "partes", es decir, bloques o "marcos", que se convierten mediante una o más entidades. Cómo se puede realizar el reparto de tareas se discute más adelante. El dispositivo (móvil) puede en el escenario mínimo sólo ocuparse del procesamiento previo de la voz digital, en cuyo caso el dispositivo externo ejecutará las etapas de análisis computacionalmente más exigentes, por ejemplo, fuerza bruta.

Correspondientemente, el intercambio de información se refiere a la interacción (recepción y/o transmisión de la información) entre el dispositivo electrónico y la entidad externa para ejecutar el proceso de conversión y procesos posteriores opcionales. Por ejemplo, la señal de voz de entrada puede ser completa o parcialmente transferida entre los por lo menos dos elementos citados anteriormente, de modo que la carga de la tarea total es compartida y/o las tareas específicas son manejadas por un elemento determinado tal como se ha mencionado en el párrafo anterior. Además, varios mensajes de parámetros, estado, reconocimiento y control pueden ser transferidos durante la etapa de intercambio de información. Otros ejemplos se describen en la descripción detallada. Los formatos de datos adecuados para la ejecución de voz o texto también se discuten.

En un aspecto adicional, se proporciona un servidor para llevar a cabo al menos una parte de la conversión de voz a texto tal como se establece en la reivindicación 6, siendo el servidor operativo en una red de comunicaciones, comprendiendo el servidor:

- una medios de entrada de datos para recibir datos digitales enviados por un dispositivo de terminal, representando dichos datos digitales la señal de voz, y uno o más comandos de control, estando cada comando asociado temporalmente con un instante de tiempo determinado en los datos digitales y determinando uno o más signos de puntuación, símbolos, u otros elementos de control que implican la manipulación de texto, y

- al menos parte de un motor de reconocimiento de voz para llevar a cabo las tareas de conversión de datos digitales a texto, en el que el motor está adaptado para posicionar físicamente, tal como en el caso de dichos signos de puntuación y símbolos, o al menos lógicamente, a través de la manipulación de texto en el caso de que dichos otros elementos de control, cada una de dichas marca de puntuación, símbolo u otro elemento de control implique la manipulación de texto en una ubicación de texto correspondiente al instante de tiempo determinado relativo a la señal de voz representada por los datos digitales recibidos, así como para cultivar el procedimiento de conversión de voz a texto al menos en parte procurado por el servidor.

El servidor también puede comprender unos medios de salida de datos para comunicar al menos parte de la salida de las tareas realizadas a una entidad externa.

Además, el servidor puede, en otro aspecto, proporcionar una ayuda especial para las personas ciegas o con problemas de visión, proporcionando funcionalidad para confirmar porciones de texto convertidas de voz a texto inciertas, según el criterio predeterminado, a través de una serie de opciones clasificadas entre sí.

En consecuencia, dicha al menos parte del motor de reconocimiento del servidor puede además ser configurada para producir un resultado de conversión de voz a texto que incluya una porción convertida, tal como una o más palabras o frases, que comprende múltiples, dos o más, opciones del resultado de la conversión, cuando la corrección del resultado de la conversión se considera como incierta para la porción de acuerdo con el criterio predeterminado, y unos medios de salida de datos para comunicar el resultado de la conversión y al menos la indicación de las opciones al terminal u otro dispositivo remoto y, opcionalmente, la activación del terminal que comprende un sintetizador de texto a voz y un medio de salida de audio, u otro dispositivo remoto, para reproducir una forma audible una o más de dichas opciones para permitir la confirmación de un resultado de la conversión deseada para la porción mediante el terminal del usuario u otro dispositivo remoto en respuesta a la reproducción audible.

La activación de la reproducción puede llevarse a cabo a través de una solicitud explícita o implícita, por ejemplo. En el caso implícito, el software del terminal está configurado para reproducir automáticamente de manera audible al menos una opción a la recepción de la misma. La petición explícita puede incluir un mensaje separado o, por ejemplo, un cierto valor de parámetro en un mensaje más genérico.

Los diversos aspectos antes mencionados de los dispositivos electrónicos y servidores se pueden combinar en un sistema que comprende al menos un dispositivo de terminal electrónico y un aparato de servidor para el reconocimiento de voz a texto cultivado. En cuanto al intercambio opcional de tareas, el sistema de conversión de voz en texto puede comprender un dispositivo terminal, por ejemplo, un terminal móvil, operable en una red de comunicaciones inalámbricas y un servidor funcionalmente conectado a la red de comunicaciones

inalámbricas, en el que el dispositivo de terminal está configurado para recibir la voz y convertir la voz en una señal de voz digital representativa, para el intercambio de información relativa a la señal digital de voz y la conversión de voz a texto de la misma con el servidor, y para ejecutar parte de las tareas necesarias para llevar a cabo una conversión de una señal voz digital a texto, y dijo servidor está configurado para recibir la información relativa a la señal digital de voz y su conversión de voz a texto, y para ejecutar, en base a la información intercambiada, la parte restante de las tareas requeridas para llevar a cabo una conversión de señal de voz digital a texto.

El "servidor" se refiere aquí a una entidad, por ejemplo, un aparato electrónico, tal como un ordenador que coopera con el dispositivo electrónico de la invención para obtener la señal de voz de origen, realizar la conversión de voz a texto, representar los resultados, o ejecutar posibles procesos adicionales. La entidad puede estar incluida en otro dispositivo, por ejemplo, una puerta de entrada o un router, o puede ser un dispositivo completamente separado o una pluralidad de dispositivos que forman la entidad del servidor agregado de la invención.

En un aspecto adicional, se proporciona por un procedimiento tal como se expone en la reivindicación 9 para la conversión de voz en texto que comprende:

- obtener una señal de voz digital y un comando de control relativo a la misma de manera de superposición temporal, en el que el comando de control determina uno o más signos de puntuación, símbolos u otros elementos de control que implican la manipulación del texto,

- asociar el comando de control con un instante de tiempo que corresponde sustancialmente a la señal de voz digital a la que se dirige el comando de control, y

- realizar una conversión de voz a texto, en la que cada marca puntuación, símbolo u otro elemento de control que implica la manipulación del texto determinado por el comando de control está físicamente, tal como por ejemplo en el caso de dichos signos de puntuación y símbolos, o al menos lógicamente, a través de la manipulación del texto en el caso de dichos otros elementos de control, situado en una ubicación del texto correspondiente a la comunicación instantánea relativa a la señal de voz para procurar el procedimiento de conversión de voz a texto.

En un aspecto adicional, el procedimiento para llevar a cabo al menos parte de un procedimiento de conversión de voz a texto mediante uno o más dispositivos electrónicos también puede comprender:

- obtener un resultado de conversión de voz a texto que incluye una porción convertida, tal como una o más palabras o frases, que comprende múltiples, dos o más, opciones de resultados de conversión,

- reproducir de manera audible una o más de dichas opciones,

- obtener una confirmación del usuario de una de dichas una o más opciones,

- seleccionar la conversión respecto a la porción convertida de acuerdo con la confirmación obtenida.

Además, los dispositivos pueden intercambiar información relativa a la señal digital de voz y la conversión de voz en texto de la misma con fines de intercambio de tareas, por ejemplo.

Sin embargo, la señal de voz digitalizada puede ser visualizada en una pantalla del terminal, de manera que las tareas de edición y confirmación también pueden estar basadas en la visualización.

La utilidad de la invención se debe a varios factores. Los comandos de control y las marcas de puntuación asociadas u otros elementos ofrecen varias ventajas. En primer lugar, el texto resultante puede ser convenientemente finalizado, ya durante el dictado se pueden omitir la silabación separada, para puntuación por ejemplo. En segundo lugar, el motor de reconocimiento de voz puede proporcionar una mayor precisión, ya que metadatos disponibles en tiempo real explícitamente indican al motor la posición sustancialmente exacta de por lo menos algunos de dichos signos de puntuación u otros elementos. Los resultados de la conversión situados antes y después de las posiciones de los metadatos pueden ser más fáciles de averiguar que la puntuación y otros puntos fijos de guía y su naturaleza, pueden proporcionar información fuente adicional para el cálculo del reconocimiento más probable y los resultados de la conversión. La función de la reproducción sonora de las opciones de conversión permite también el análisis auditivo y la verificación de los resultados de la conversión, además o en lugar de la mera verificación visual. Este es un beneficio particular para las personas ciegas o dificultades de visión, que también pueden estar interesadas en la utilización las tareas de conversión de voz a

texto. Además, las personas con buena visión pueden aprovechar la función de verificación audible cuando prefieren utilizar su visión para otros fines.

Con la ayuda de varios ejemplos de la presente invención, uno puede generar una forma textual de mensajes para propósitos de archivo y/o comunicaciones con facilidad hablando a su dispositivo electrónico, posiblemente móvil, y opcionalmente editando la señal de voz a través de la interfaz de usuario mientras el dispositivo y la entidad conectada de forma remota automáticamente se cuida de la conversión exhaustiva de voz a texto. La práctica de la comunicación entre el dispositivo móvil y la entidad puede soportar una pluralidad de diferentes medios (llamadas de voz, mensajes de texto, protocolos de transferencia de datos móviles, etc.) y la selección de un procedimiento de intercambio de información puede hacerse incluso de forma dinámica sobre la base de las condiciones de la red, por ejemplo. El texto resultante y/o la voz editada se puede comunicar hacia adelante a un destinatario predeterminado mediante la utilización de una pluralidad de diferentes tecnologías y técnicas de comunicación, incluyendo Internet y redes móviles, intranets, correo de voz (síntesis de voz requerido para el texto resultante), correo electrónico, mensajes SMS/MMS, etc. Texto como tal puede ser proporcionado en forma editable o de sólo lectura. Formatos de texto aplicables incluyen ASCII plano (y otros conjuntos de caracteres), formato MS Word, y formato Adobe Acrobat, por ejemplo.

El dispositivo electrónico de los diversos ejemplos de la presente invención puede ser un dispositivo o estar al menos incorporado en un dispositivo que el usuario lleva con él, en cualquier caso, y así no se introduce carga adicional. Como el texto puede ser sometido adicionalmente a un motor de traducción automática, la invención también facilita la comunicación multilingüe. Siempre que la capacidad de edición manual de la señal de voz permita al usuario verificar y cultivar la señal de voz antes de la ejecución de nuevas acciones, que pueden gastar el sistema con un procesamiento innecesario y, en ocasiones, mejorar la calidad de la conversión cuando el usuario puede reconocer por ejemplo, porciones inarticulados en el grabado de la señal de voz y reemplazarlos con versiones apropiadas. El posible reparto de tareas entre el dispositivo electrónico y la entidad externa puede ser configurable y/o dinámico, lo que aumenta considerablemente la flexibilidad de la solución global, como la transmisión de datos disponibles y el procesamiento de memoria/recursos, sin olvidar otros aspectos como el consumo de la batería, el servicio de fijación de precios/contratos, preferencias del usuario, etc. se pueden tener en cuenta incluso en tiempo real en la explotación de la invención, el dispositivo móvil y el usuario específicamente. El aspecto de la personalización de la parte de reconocimiento de voz de la invención, respectivamente, aumenta la calidad de la conversión.

El núcleo de la actual invención puede ser convenientemente expandido a través de servicios adicionales. Por ejemplo, servicios de revisión ortográfica manual/automática o de verificación de traducción de idiomas y/o traducción pueden ser introducidos en el texto, ya sea directamente por el operador del servidor o por un tercero al que el dispositivo móvil y/o el servidor transmiten los resultados de la conversión. Además, el lado del servidor de la invención puede ser actualizado con lo último en hardware/software (por ejemplo, software de reconocimiento) sin necesidad de plantear una necesidad de la actualización del dispositivo(s) electrónico(s), como móviles. En consecuencia, el software puede ser actualizado a través de la comunicación entre el dispositivo y el servidor. Desde un punto de vista de servicio, dicha interacción abre nuevas posibilidades para la definición de una jerarquía de nivel de servicio integral. Como por ejemplo, dispositivos móviles, terminales móviles, por ejemplo, típicamente tienen capacidades diferentes y los usuarios de los mismos son capaces de gastar una suma variable de dinero (por ejemplo, en una forma de costes de transferencia de datos o tarifas de servicio directo) para la utilización de la invención, diversas versiones de software móvil pueden estar disponibles; la diferenciación se puede implementar a través de características de bloqueo/activación o aplicaciones totalmente independientes para cada nivel de servicio. Por ejemplo, en un nivel, las entidades de red se encargarán de la mayoría de las tareas de conversión y el usuario está dispuesto a pagar por las mismas, mientras que en otro nivel el dispositivo móvil ejecuta una parte sustantiva del proceso, ya que tiene las capacidades necesarias y/o el usuario no quiere utilizar los recursos externos para ahorrar costes, o por alguna otra razón.

En un ejemplo de la invención, una disposición de conversión de voz a texto después de los principios explicados anteriormente se aplican de tal manera que una persona utilizada para dictar memos utiliza su dispositivo de computación de usos múltiples para la captura de una señal de voz en cooperación con la característica simultánea basada en comandos de control, edición/sección. En otro ejemplo, ya sea independiente o complementario, la reproducción audible de las opciones de resultado de la conversión se aprovecha para facilitar la determinación del resultado de la conversión final. Las variaciones de los ejemplos también se describen.

Breve descripción de los dibujos

A continuación, la invención se describe con más detalle con referencia a los dibujos adjuntos, en los que

La figura 1 ilustra un diagrama de flujo de un escenario de la técnica anterior relativa al software de reconocimiento de voz.

La figura 2a ilustra un ejemplo de la presente invención, en el que uno o más comandos de control se proporcionan durante el procedimiento de grabación de voz para cultivar la conversión de voz a texto.

La figura 2b ilustra un escenario, que puede cooperar con el ejemplo de la figura 2a o utilizarse de forma independiente, en el que se proporcionan múltiples opciones de conversión de voz a texto y una o más de las mismas se reproducen de manera audible para la obtención de la confirmación de la opción deseada.

La figura 2c visualiza el intercambio de comunicación y/o tareas entre varios dispositivos durante el procedimiento de conversión de voz a texto.

La figura 3a revela un diagrama de flujo de una opción para llevar a cabo un primer ejemplo del procedimiento de la presente invención.

La figura 3b revela otro diagrama de flujo para llevar a cabo un segundo ejemplo, ya sea independiente o complementario, del procedimiento de acuerdo con la presente invención.

La figura 3c revela un diagrama de flujo relativo a la edición de la señal y teniendo lugar el intercambio de datos potencialmente en el contexto de la presente invención.

La figura 4 revela un gráfico de señalización que muestra las posibilidades de transferencia de información entre los dispositivos para la aplicación de un ejemplo deseado de la invención actual.

La figura 5 representa una realización, meramente de ejemplo, de los componentes internos del motor de reconocimiento de voz con una serie de tareas.

La figura 6 es un diagrama de bloques de un ejemplo de un dispositivo electrónico de la presente invención.

La figura 7 es un diagrama de bloques de un ejemplo de una entidad de servidor de acuerdo con la presente invención.

Descripción detallada de las realizaciones

La figura 1 ya fue revisada en conjunción con la descripción de la técnica anterior relacionada.

La figura 2a revela un ejemplo de la presente invención en el que un comando de control se proporciona durante el procedimiento de grabación de voz para cultivar la conversión de voz a texto en relación particularmente al instante de voz y la posición de texto correspondiente para que se le dio el comando.

El dispositivo electrónico 202 puede ser un terminal móvil, una PDA, una máquina de dictado, o un ordenador de sobremesa o portátil, por ejemplo. Hay dos opciones, a saber, un terminal móvil y un ordenador portátil, que se ilustran de forma explícita en la figura. El dispositivo 202 está provisto de medios que incluyen hardware y software (lógica) para introducir voz. Los medios pueden incluir un micrófono para recibir una señal acústica y un convertidor A/D para su conversión en forma digital. Por otra parte, los medios sólo pueden recibir una señal de audio de forma digital ya capturada desde un dispositivo remoto, tal como un micrófono inalámbrico o con cable. Además, el dispositivo comprende unos medios de control de entrada integrados, o al menos funcionalmente conectados, tales como un teclado, botón(es), mando(s), corredera(s), control remoto, controlador de voz (incorporando micrófono y software de interpretación, por ejemplo), o por ejemplo, una pantalla táctil para introducir un comando de control simultáneamente con la obtención de la señal de voz digital. El dispositivo 202, por lo tanto, monitoriza uno o más comandos de control, similares o diferentes del usuario del dispositivo, mientras se obtiene la señal de voz digital. El dispositivo 202 está configurado para asociar temporalmente el comando de control con un instante de tiempo que corresponde sustancialmente a la señal de voz digital, sobre la cual se comunica el comando de control. Esta asociación puede realizarse por software de dictado u otro software que se ejecuta en el dispositivo 202.

Los medios de entrada de control pueden comprender una pluralidad de elementos de entrada, tales como claves diferentes que pueden estar asociadas, por ejemplo, a través de software, con diferentes elementos de control, preferiblemente definibles por el usuario, tales como signos de puntuación u otros elementos, opcionalmente simbólicos, indicados por los comandos de control para cultivar el procedimiento de conversión

de voz a texto. Un elemento de entrada puede estar asociado con al menos un elemento de control, pero por ejemplo, una activación rápida múltiple del mismo elemento de entrada puede implicar también, a través de un comando específico o dos comandos similares temporalmente adyacentes, un elemento de control diferente al de una activación más alejada. El elemento de control puede incluir distintas marcas de puntuación y otros símbolos, incluyendo pero no limitado a, cualquier elemento seleccionado de un grupo formado por: dos puntos, coma, guión, apóstrofe, corchete (por ejemplo, corchetes u otros elementos vinculados, el mismo elemento de entrada puede Inicialmente, en primera instancia de activación, referirse a una apertura del soporte/elemento y, a continuación, en el ejemplo siguiente, a un cierre del soporte/elemento, o a la apertura y cierre de paréntesis o elementos que pueden ser asignados a elementos de entrada diferentes), elipsis, signo de exclamación, punto, comillas, guión, signo de interrogación, punto y coma, barra, signo de número, símbolo de moneda, signo de sección, asterisco, barra invertida, avance de línea, y espacio. Así, los elementos de control se pueden introducir como tales en el texto convertido, y/o pueden implicar realizar alguna manipulación del texto (por ejemplo, la inserción de espacios o filas, una letra grande inicial, eliminar una sección predeterminada anterior, por ejemplo, hasta un elemento anterior, tales como un periodo, etc.) en la posición asociada. Por lo tanto, se puede decir que los elementos están al menos lógicamente posicionados en un lugar de texto correspondiente a la comunicación instantánea relativa a la señal de voz digital, así como para cultivar el procedimiento de conversión de voz a texto.

Los elementos de control pueden facilitar el proceso de reconocimiento de voz como por ejemplo, la probabilidad de la existencia de una cierta formulación predeterminada alrededor de un elemento de control predeterminado, tal como una marca de puntuación (es decir, el contexto), puede ser generalmente mayor que la probabilidad de la existencia de otras formulaciones en relación con ese elemento en particular, y si uno o más resultados del reconocimiento locales son por el contrario inciertos debido al hecho de que la señal de entrada igualmente coincide con varias opciones de reconocimiento diferentes, el comando de control puede definir un elemento tal como una marca de puntuación que afecta las probabilidades y por lo tanto potencialmente facilita la selección del resultado del reconocimiento más probable respecto al texto anterior, siguiente, o circundante.

Además, al menos algunos de los comandos pueden estar asociados con las operaciones suplementarias, tales como una pausa de grabación de longitud predeterminada. La pausa (comienzo y/o final) u otra función pueden ser indicada al usuario del dispositivo 202 por un signo visual (pantalla), táctil (por ejemplo vibración) o de audio (a través de un altavoz), por ejemplo. Por ejemplo, la entrada asociada a un punto o una coma también podría estar vinculada con una pausa para que el usuario pueda continuar dictando de forma natural y ordenar sus pensamientos para la siguiente frase, etc. Preferiblemente, el usuario puede configurar las asociaciones entre los diferentes comandos, elementos de entrada, y/o operaciones complementarias.

El dispositivo 202 puede grabar la voz y datos de comandos de control asociados primero a nivel local, o tampón en tiempo real y enviarlo a un servidor remoto 208 que se puede conectar al dispositivo 202 a través de una o más redes de comunicaciones inalámbricas 204 y/o por cable 206. En el primer caso, el dispositivo 202 puede, después de adquirir todos los datos, pasarlos hacia adelante para el reconocimiento de voz remoto y la conversión de voz a texto.

Alternativamente, el dispositivo 202 puede comprender todos los medios necesarios para realizar localmente la conversión de voz a texto, que se ilustra mediante el rectángulo 220 mientras los elementos externos/remotos 204, 206, y 208 se ilustran en una forma de rectángulo vecino. En una alternativa adicional, el reparto de operaciones entre los dispositivos local y remoto se puede aplicar para ser revisada con más detalle más adelante en este documento. El número de referencia 212 implica la transferencia de datos, por ejemplo, salida de conversión de resultado, a entidades externas adicionales.

Normalmente, las redes inalámbricas comprenden transceptores de radio llamados por ejemplo a las estaciones base, o puntos de acceso para la conexión de los dispositivos terminales. La comunicación inalámbrica puede también referirse al intercambio de otros tipos de señales que sólo señales de frecuencia de radio, incluyendo dichos otros tipos de señales, por ejemplo las señales de infrarrojos, o de ultrasonido. Operativamente en alguna red se refiere aquí a la capacidad de transferencia de información.

La red de comunicaciones inalámbrica 204 puede estar conectada además a otras redes, por ejemplo, una red de comunicaciones (por cable) 206, a través de medios de interfaz apropiados, por ejemplo, routers o interruptores. Conceptualmente, por ejemplo, la red inalámbrica 204 también se puede encontrar directamente en la red de comunicaciones 206 si se trata de nada más que una interfaz inalámbrica para comunicarse con las terminales inalámbricas al alcance. Un ejemplo de la red de comunicaciones 206, que también abarca una pluralidad de subredes tal como Internet.

En el caso de que una o más entidades externas tales como el servidor 208 se ocupen de al menos parte del proceso general, diferentes actividades de transferencia de datos pueden tener lugar desde/hacia el dispositivo 202, como se ilustra mediante la flecha bidireccional discontinua. Por ejemplo, los datos de voz digitales, datos de comando de control (cc), y el texto convertido pueden ser transferidos.

En 222 se presenta una ilustración de un procedimiento de conversión de voz a texto cultivada por el procedimiento de adquisición de comandos de control en tiempo real. Una forma ondulada ilustra un discurso señal de audio grabada que comprende voz y la línea vertical discontinua 224 indica un instante de tiempo en que el usuario del dispositivo 202 proporcionan un comando de control asociado con un período u otro elemento que se coloca en el lugar correspondiente en el resultado de la conversión. Una ilustración como tal también se puede proporcionar en una pantalla del dispositivo 202, si se desea.

En lugar de o además de adquisición de datos de comandos de control mientras se obtiene la señal de voz, los comandos de control pueden ser grabados después durante la reproducción de una señal de audio ya grabada, por ejemplo.

La figura 2b describe un escenario que puede ser integrado con el ejemplo de la figura 2a, o aplicado como una solución independiente. La transferencia de datos entre diferentes entidades en general, puede tener lugar como en el ejemplo anterior o el dispositivo 202 puede volver a ser completamente autónomo con respecto a las operaciones realizadas.

En el ejemplo ilustrado, el dispositivo 202, el servidor 208, o una combinación de varias entidades tales como el dispositivo 202 y el servidor 208, han procesado la señal de voz de entrada de tal manera que se ha obtenido un texto resultado de la conversión 226 con una o más porciones convertidas que se extienden desde un solo símbolo o palabra a una frase, por ejemplo, cada una de ellas incluyendo múltiples, es decir, dos o más, opciones de resultado de la conversión. Las opciones son preferentemente representadas al usuario para su revisión y selección/confirmación en un orden predeterminado, por ejemplo, opción más probable primero. Las opciones son preferentemente reproducidas audiblemente a través de tecnología TTS (texto a voz) y, por ejemplo uno o más altavoces, mediante el dispositivo 202, pero alternativamente o adicionalmente, también pueden ser utilizadas la reproducción visual o táctil por ejemplo. En la reproducción visual, las opciones pueden mostrarse como una secuencia o una lista (horizontal o vertical) en una pantalla, una o más opciones a la vez. En el caso de múltiples opciones mostradas simultáneamente la actualmente seleccionada puede mostrarse como destacada. En la reproducción táctil, por ejemplo, un elemento/unidad de vibración acoplada al dispositivo 202 puede señalar las opciones mediante un código bien definido tal como el código Morse. Véase por ejemplo la ilustración 228 en el rectángulo 226 (una vista de la pantalla, por ejemplo) que representa una porción del resultado de la conversión indicado por las líneas discontinuas y teniendo tres opciones probables, seleccionadas de acuerdo con un criterio predeterminado.

En un ejemplo, las opciones reales y señales de guía opcionales (por ejemplo solicitud para seleccionar la opción deseada accionando un elemento de entrada de IU predeterminado) pueden ser reproducidas en forma audible sobre toda la reproducción del resultado de la conversión global, es decir, el dispositivo 202 puede ser configurado para reproducir audiblemente el resultado de la conversión entero como un documento dictado, y para pedir al usuario sobre la instancia de cada una de las porciones antes mencionadas qué opción se debe elegir como la porción convertida final.

En otro ejemplo, al menos las porciones antes mencionadas y opcionalmente las señales de guía se reproducirán al usuario para la selección.

En un ejemplo, la opción más probable, se reproduce primero de tal manera que si el usuario está contento con ella, él/ella puede inmediatamente aceptarla y ahorrar algo de tiempo de la revisión de las otras opciones inferiores.

Los medios de entrada de control nuevamente pueden comprender llaves, mandos, etc., como ya se ha revisado en relación con el ejemplo de la figura 2a.

Después de obtener la selección del usuario de la opción deseada para una o más porciones antes mencionadas, las opciones descartadas pueden ser eliminadas y la seleccionada ser incrustada en el resultado de la conversión final.

La figura 2c describe un bosquejo de un sistema, solamente a modo de ejemplo, adaptado para llevar a cabo un escenario de la disposición de conversión de la invención como se ha descrito anteriormente bajo el control de un usuario que favorece grabar sus mensajes y conversaciones en lugar de escribirlos en su multipropósito

móvil u otro dispositivo electrónico que proporciona una UI al resto del sistema. Uno o más características de este escenario puede ser combinada con las características de los ejemplos de la figura 2a y/o la figura 2b. El dispositivo electrónico 202, como el terminal móvil o un PDA con un medio de comunicación interno o externo, por ejemplo, un transceptor de radio frecuencia, es capaz de funcionar en una red inalámbrica de comunicaciones 204 como una red celular o red WLAN (LAN inalámbrica), capaz de intercambiar información con el dispositivo 202.

El dispositivo 202 y el servidor 208 intercambian información 210 a través de redes 204, 206 con el fin de llevar a cabo el proceso de conversión general de voz a texto. Un motor de reconocimiento de voz se encuentra en el servidor 208 y, opcionalmente, al menos parcialmente en el dispositivo 202. El texto resultante y/o la voz editada pueden ser comunicada 212 luego hacia un destinatario remoto dentro o fuera de dichas comunicaciones inalámbricas 204 y redes de comunicaciones 206, un archivo electrónico (en cualquier red o en el dispositivo 202, por ejemplo, en una tarjeta de memoria), o una entidad de servicio que se encargue del procesamiento adicional, traducción, por ejemplo, de la misma. El tratamiento posterior, alternativamente/adicionalmente, puede llevarse a cabo en el servidor 208.

En un ejemplo complementario o independiente de la presente invención, un usuario puede estar dispuesto a incrustar nuevos datos de voz o textuales en una muestra de voz existente (por ejemplo un archivo) o texto convertido a partir de la misma, respectivamente. Por ejemplo, el usuario dicta, por ejemplo, una cantidad de 30 minutos de voz, pero luego se da cuenta que quiere decir algo más, ya sea a) que puede caer en medio de otros dos archivos de sonido grabados con anterioridad o b) en un archivo de sonido existente. El dispositivo 202 y/o el servidor 208 puede entonces ser configurado para incrustar los nuevos datos de voz en la muestra de voz existente directamente o a través de metadatos (por ejemplo, a través de un archivo de enlace que temporalmente asocia una pluralidad de archivos de muestra de voz) para la posterior conversión de todos los datos de voz en uno o más archivos. En el caso de que la porción original de 30 minutos ya ha sido convertida en texto, el usuario sólo puede definir en el archivo de audio de origen y/o el archivo de texto resultante a través de la UI un lugar apropiado para la nueva porción de voz y el texto correspondiente de tal manera que sólo la nueva porción de voz pueda ser entonces convertida en texto e incrustada en el resultado de la conversión ya disponible. Como un ejemplo de aplicación, el usuario puede escuchar o desplazarse visualmente a través de la voz original y/o el texto resultante y determinar una posición para insertar el tipo de la nueva grabación que se realizará entonces, con lo cual el dispositivo 202 y/o el servidor remoto 208 se ocupan de los procedimientos restantes, tales como la conversión de voz a texto, transferencia de datos, o la integración de los resultados de la conversión.

Volviendo de nuevo a la figura 2c, los bloques 214, 216 representan instantáneas de vista potencial de la pantalla del dispositivo 202 tomadas sobre la ejecución del procedimiento de conversión de texto a voz en su conjunto. La instantánea 214 ilustra una opción para visualizar, mediante una aplicación de conversión, la señal de entrada (es decir, la señal de entrada que comprende al menos voz) al usuario del dispositivo 202. La señal puede de hecho ser visualizada para su revisión y edición mediante la capitalización de una serie de enfoques diferentes: la representación en el dominio temporal de la señal puede ser dibujada como un sobre (ver la curva superior en la instantánea) o como un gráfico más grueso (por ejemplo, de tipo voz encendido/apagado o de otra segmentación de dominio temporal de resolución reducida, en cuyo caso la resolución reducida puede ser obtenida a partir de la señal original dividiendo el rango de valor original del mismo en un número menor de subrangos limitados a un valor umbral, por ejemplo) basado en el los valores de amplitud o magnitud de los mismos, y/o un espectro de potencia u otra parametrización de dominio de frecuencia/altemativo puede calcularse a partir del mismo (véase la curva inferior en la instantánea).

Varias técnicas de visualización pueden aplicarse incluso simultáneamente, en donde a través de por ejemplo un zoom (/deshacer el zoom) o alguna otra funcionalidad una cierta porción de la señal correspondiente a un intervalo de tiempo definido por el usuario o un sub-rango de valores de parámetros preferidos puede ser mostrado en otro lugar en la pantalla (ver las curvas superior e inferior de la instantánea 214 presentadas simultáneamente) con la resolución aumentada(/disminuida) o mediante una técnica de representación alternativa. Además de la representación(es) de la señal, la instantánea 214 muestra varios valores numéricos determinados durante el análisis de la señal, marcadores (rectángulo) y puntero (flecha, línea vertical) a la señal (porción), y funciones actuales de edición o de visualización de datos aplicadas o disponibles, consulte el número de referencia 218. En el caso de una pantalla sensible al tacto, el usuario puede ventajosamente pintar con el dedo o el lápiz una zona preferente de la porción de la señal visualizada (la señal puede ventajosamente ser desplazada por el usuario si por lo contrario no cabe en la pantalla con una resolución preferida) y/o pulsando otra, área predeterminada, especifica una función que se ejecuta en relación con la porción de señal subyacente en la zona preferida. Una funcionalidad similar se puede proporcionar al usuario a través de medios de control más convencionales, por ejemplo, un puntero en movimiento en la pantalla en respuesta a la señal de

entrada de control del dispositivo creada por un controlador de punto de seguimiento, un ratón, un botón del teclado/teclado, un controlador direccional, un receptor de mando de voz, etc.

A partir de la señal visualizada, el usuario del dispositivo 202 puede reconocer rápidamente, con el único requisito de una experiencia menor, las expresiones separables tales como palabras y objetos posibles (ruidos de fondo, etc.) contenidos en el mismo y además editar la señal con el fin de cultivarla para el posterior proceso de reconocimiento de voz. Si por ejemplo se muestra un sobre de la representación del dominio de tiempo de la señal de voz, porciones de amplitud inferior a lo largo del eje de tiempo corresponden, con una alta probabilidad, al silencio o al ruido de fondo mientras que las expresiones de voz contienen más energía. En el dominio de la frecuencia de los picos dominantes son, respectivamente, debidos a los componentes de la señal real de voz.

El usuario puede entrar y comunicar comandos de edición de señal al dispositivo 202 a través de la IU de los mismos. Funciones de edición de señal relacionadas con los comandos de preferencia deberán permitir la inspección y la revisión integral de la señal original, revelándose a continuación unos pocos ejemplos útiles.

La porción de la señal definida por el usuario (por ejemplo, ya sea seleccionado con marcadores/punteros movibles o "pintados" en la IU tales como la pantalla táctil como se explicó anteriormente) será reemplazable con otra porción, ya almacenada o bien grabada en tiempo real. Asimismo, una porción será eliminable de manera que las porciones adyacentes restantes se unan entre sí o la porción eliminada se sustituye con algunos datos predeterminados que representan por ejemplo el silencio o ruido de fondo de bajo nivel. En los extremos de la señal captada dicho procedimiento de unión no es necesario. El usuario puede ser asignado con una posibilidad de alterar, por ejemplo unificar, la amplitud (en relación volumen/ruido) y el contenido espectral de la señal, que puede llevarse a cabo a través de diferentes medios de control de ganancia, algoritmos de normalización, un ecualizador, un controlador de rango dinámico (incluyendo por ejemplo una puerta de ruido, expansor, compresor, limitador), etc. Los algoritmos de reducción de ruido para el esclarecimiento de la señal de voz degradado del ruido de fondo son más complejos que el ruido, pero ventajosos siempre que la señal acústica original se ha producido en condiciones de ruido. El ruido de fondo de preferencia será de al menos pseudo-estacionario para garantizar la exactitud de modelado adecuada. Los algoritmos modelan el ruido de fondo espectralmente o por medio de un filtro (coeficientes) y restan la estimación del ruido modelada a partir de la señal del micrófono capturada en el dominio temporal o espectral. En algunas soluciones la estimación del ruido sólo se actualiza cuando un detector de actividad de voz independiente (VAD) notifica que no hay voz en la porción de señal actualmente analizada. La señal generalmente se puede clasificar como incluyendo el sólo ruido, sólo voz, o ruido + voz.

La aplicación de conversión puede almacenar un número de diferentes funciones de edición de señales y algoritmos que son seleccionables por el usuario como tales, y al menos algunos de ellos pueden estar además adaptados por el usuario por ejemplo a través de un número de parámetros ajustables.

La funcionalidad cancelar, también conocida como funcionalidad "deshacer", siendo por ejemplo un conmutador de programas para volver a la condición de la señal antes de la última operación, se incluye preferiblemente en la aplicación a fin de permitir al usuario experimentar de forma segura con los efectos de diferentes funcionalidades mientras que busca una señal editada óptima.

Siempre que la edición se produce, al menos parcialmente de forma simultánea con el reconocimiento de voz, incluso sólo el texto así resultante puede ser visualizado en la pantalla del dispositivo 202. Esto puede requerir la transferencia de información entre el servidor 208 y el dispositivo 202, si el servidor 208 ha participado en la conversión de la porción de voz particular, a partir de la se ha originado que el texto así resultante. De lo contrario, la instantánea 216 se materializa después de terminar la conversión de voz a texto. Alternativamente, el texto como tal, nunca se muestra al usuario del dispositivo 202, al ser, por defecto, directamente transferido hacia adelante hasta el destino de archivo o un receptor remoto, preferiblemente dependiendo de los ajustes definidos por el usuario.

Una configuración puede determinar si el texto se muestra automáticamente en la pantalla del dispositivo 202 para revisión, de nuevo, opcionalmente junto con la señal de voz original o editada, es decir, la señal de voz se visualiza como se ha descrito anteriormente, mientras que las porciones de texto resultantes tales como palabras se muestran por encima o por debajo de la voz como estando alineadas en relación a las porciones de voz correspondientes. Los datos necesarios para la alineación se crean como un subproducto en el proceso de reconocimiento de voz en el que la señal de voz ya se analizó en porciones. El usuario puede determinar entonces si él está contento con el resultado de la conversión o decidir seguir editando las porciones preferidas de voz (incluso volver a grabar las mismas) y someterlas a una nueva ronda de reconocimiento, manteniendo intactas las porciones restantes, si las hubiere. Este tipo de conversión recursiva de voz a texto es cierto que consume más tiempo y recursos que el enfoque más directo de tipo básico "editar de una vez y convertir", pero

permite resultados más precisos que deben alcanzarse. Alternativamente, al menos parte del texto resultante puede ser corregido insertando correcciones manualmente con el fin de omitir rondas adicionales de conversión sin verdadera certeza de obtener resultados más precisos.

Aunque la señal de audio de entrada que comprende la voz es originalmente capturada por el dispositivo 202 a través de un sensor o transductor tal como un micrófono y luego digitalizada a través de un convertidor A/D para la transmisión y/o almacenamiento de forma digital, incluso la fase de edición puede comprender transferencia de información entre el dispositivo 202 y otras entidades, tales como el servidor 208 según lo previsto por el enfoque recursivo anterior. Respectivamente, la señal de voz digital puede ser tan grande en tamaño que no puede ser sensiblemente almacenada en el dispositivo 202, como tal, por lo que tiene que ser comprimido localmente, opcionalmente en tiempo real durante la captura, utilizando una voz dedicada o codificador de audio más genérico tal como GSM, TETRA, G.711, G.721, G.726, G.728, G.729, o varios codificadores de series MPEG. Además, o alternativamente, la señal de voz digital puede, durante la captura, ser transmitida directamente (incluyendo sin embargo la memoria intermedia necesaria) a una entidad externa, por ejemplo, el servidor 208, para el almacenamiento y opcionalmente la codificación, y posteriormente se recupera de nuevo al dispositivo 202 para su edición. En el caso extremo, la edición tiene lugar en el servidor 208 de tal manera que el dispositivo 202 principalmente actúa como una interfaz remota para controlar la ejecución de las funciones de edición antes explicadas en el servidor 208. Para tal fin, tanto datos de voz (para la visualización en el dispositivo 202) e información de control (comandos de edición) tienen que ser transferidos entre las dos entidades 202, 208.

El intercambio de información 210 como un conjunto puede incorporar una pluralidad de diferentes características de la disposición de conversión. En un aspecto de la invención, el dispositivo 202 y el servidor 208 comparten las operaciones relacionadas con la conversión de voz a texto. El reparto de operaciones inherentemente implica también el intercambio de información 210 ya que al menos parte de la voz (opcionalmente codificada) tiene que ser transferida entre el dispositivo 202 y el servidor 208.

Las aplicaciones de conversión en el dispositivo 202 y, opcionalmente, en el servidor 208 incluyen o tienen al menos acceso a la configuración de la operación (por ejemplo, la función, el algoritmo) compartiendo con un número de parámetros, que pueden ser definidos por el usuario o fijos (o al menos no libremente modificables por el usuario). Los parámetros pueden ya sea determinar de forma explícita cómo se dividen las operaciones entre el dispositivo 202 y el servidor 208, o sólo supervisar el proceso por una serie de normas más genéricas que deben seguirse. Por ejemplo ciertas operaciones pueden realizarse siempre mediante el dispositivo 202 o por el servidor 208. Las reglas pueden especificar compartir la carga de procesamiento, en el que los umbrales de carga relativos o absolutos con adaptabilidad/lógica opcional adicional son determinados para las cargas, tanto del dispositivo 202 y del servidor 208 para transferir por lo general parte del procesamiento y por lo tanto los datos de origen de la entidad más cargada a la menos cargada. Si el proceso de conversión de voz a texto se implementa como un servicio basado en suscripción que incluye un número de niveles de servicio, algunas de las características de conversión pueden ser desactivadas en un nivel de usuario (inferior) determinado por el bloqueo en la aplicación de conversión, por ejemplo. La funcionalidad bloqueo/desbloqueo puede llevarse a cabo a través de un conjunto de diferentes versiones de software, códigos de registro de características, módulos de software descargables adicionales, etc. En el caso de que el servidor 208 no pueda poner en práctica algunas de las operaciones de nivel más bajo permitido solicitados por el dispositivo 202, por ejemplo durante una sobrecarga del servidor o situación de servidor caído, puede enviar un mensaje "no reconocimiento" u omitir completamente el envío de cualquier respuesta (a menudo los reconocimientos son de hecho enviados como se presenta en la figura 4) de modo que el dispositivo 202 puede deducir a partir de la negativa o falta de reconocimiento ejecutar las operaciones por sí mismo cuando sea posible.

El dispositivo 202 y el servidor 208 pueden negociar un escenario de cooperación para compartir la operación y el intercambio de información resultante 210. Tales negociaciones pueden ser activadas por el usuario (es decir, seleccionando una acción conducente al inicio de las negociaciones), en una forma temporizada (una vez al día, etc.), al comienzo de cada conversión, o dinámicamente durante el proceso de conversión mediante la transmisión de información del parámetro entre sí en relación con un cambio de valor de parámetro, por ejemplo. Los parámetros relacionados con el reparto de operaciones incluye información sobre, por ejemplo uno o más de los siguientes: el tratamiento actual o la carga de memoria, estado de la batería o su capacidad máxima, el número de otras operaciones que se ejecutan (con una prioridad más alta), el ancho de banda de transmisión disponible, los costos relacionados con aspectos tales como la velocidad actual de transmisión de datos por la vía(s) de transferencia disponible o coste por el uso del servidor por tamaño/duración de voz, el tamaño/duración de la señal de voz fuente, los procedimientos de codificación/decodificación disponibles, etc.

El servidor 208 es en la mayoría de los casos superior al dispositivo 202 en cuanto a la potencia de procesamiento y capacidad de memoria, por lo tanto, las comparaciones de carga serán relativas o escaladas

de otra manera. La lógica para llevar a cabo el reparto de operaciones se puede basar en simples tablas de valores de umbral, por ejemplo, que incluyen diferentes rangos de los valores parámetros y decisiones resultantes de reparto de operaciones. La negociación podrá, en la práctica, realizarse a través de intercambio de información 210 de forma que el dispositivo 202 o el servidor 208 transmite la información de estado a la otra parte lo que determina un escenario optimizado de cooperación y señala de nuevo el resultado del análisis para iniciar el proceso de conversión.

El intercambio de información 210 también incluye la transmisión de estado de la conversión (operación actual terminada/seguimiento de ejecución de anuncios, servicio de anuncios, cifras del servicio de carga, etc.) y mensajes de señalización de reconocimiento (recepción de datos de forma exitosa/no exitosa, etc.) entre el dispositivo 202 y el servidor 208. Siempre que las asignaciones del reparto de operaciones se fijan, la transferencia de la señalización relacionada no es en cambio obligatoria.

El intercambio de información 210 puede tener lugar durante las prácticas de comunicación diferentes, incluso los múltiples de forma simultánea (transferencia de datos paralela) para acelerar las cosas. En un ejemplo, el dispositivo 202 establece una llamada de voz al servidor 208 sobre el cual se transmite la señal de voz o al menos parte de ella. La voz puede ser transferida en relación con la fase de captura, o después de la primera edición que en el dispositivo 202. En otro ejemplo, el protocolo específico de transferencia de datos tales como GPRS se utiliza para la transferencia de voz y de otra información. La información puede ser encapsulada en varios formatos de paquetes/marco de datos y mensajes como SMS, MMS o mensajes de correo electrónico.

Los resultados intermedios proporcionados por el dispositivo 202 y el servidor 208, por ejemplo voz procesada, parámetros de reconocimiento de voz, o porciones de texto, se pueden combinar en cualquiera de dichos dos dispositivos 202, 208 para crear el texto final. Dependiendo de la naturaleza de la distribución (los resultados intermedios representan las correspondientes porciones finales de texto) los resultados intermedios pueden ser alternativamente transmitidos como tales a otra entidad de recepción que puede realizar el proceso de combinación final mediante la aplicación de la información proporcionada al mismo por las entidades 202, 208 para ese fin.

Servicios adicionales tales como la corrección ortográfica, traducción de máquina/humana, verificación de la traducción o texto adicional para la síntesis de voz (TTS) pueden estar situados en el servidor 208 u otra entidad remota a la que se transmite el texto después de terminar la conversión de voz a texto. En el caso de que los resultados intermedios antes mencionados se refieran directamente a las porciones de texto, las porciones pueden ser transmitidas de forma independiente inmediatamente después de su finalización, siempre que la información adicional correspondiente a combinar también se transmita finalmente.

En una implementación de la invención, el motor de reconocimiento de voz de la invención que reside en el servidor 208 y, opcionalmente, en el dispositivo 202 puede personalizarse para utilizar las características individuales de la voz de cada usuario. Esto indica la introducción de características a una base de datos local o remota accesible por el motor de reconocimiento sobre la base por ejemplo, del ID de usuario; las características pueden obtenerse convenientemente a través de la capacitación del motor, proporcionando muestras de la voz elegidas libremente/pares de texto correspondientes para el motor o pronunciando las expresiones que el motor está configurado para solicitar a cada usuario en función de, por ejemplo, un compromiso predefinido (dependiente del idioma) entre maximizar la versatilidad y el valor de representación del espacio de información y reducir al mínimo el tamaño del mismo. Basado en el análisis de los datos de entrenamiento, el motor determina entonces la configuración personalizada, por ejemplo los parámetros de reconocimiento, a ser utilizados en el reconocimiento. Opcionalmente, el motor ha sido adaptado para actualizar continuamente la información del usuario (~perfiles de usuario) mediante la utilización de la información obtenida; las diferencias entre el texto final corregido por el usuario y el texto automáticamente producido puede ser analizado.

La figura 3a describe, a modo de ejemplo, un diagrama de flujo de un procedimiento de acuerdo con el ejemplo de la figura 2a. Durante la etapa de puesta en marcha del procedimiento 302 diversas acciones iniciales que permiten la ejecución de las etapas procedimiento pueden realizarse. Por ejemplo, las aplicaciones necesarias, una o más, en relación con el proceso de conversión de voz a texto pueden lanzarse en el dispositivo 202, y el servicio respectivo puede ser activado en el lado del servidor 208, si lo hubiere. Si el usuario del dispositivo 202 desea un reconocimiento personalizado, la etapa 302 opcionalmente, incluye el registro o el inicio de sesión en la aplicación y/o servicio asociado. Esto también se lleva a cabo siempre que el servicio esté dirigido a usuarios registrados (servicio privado) y/u ofrece una pluralidad de niveles de servicio diferentes. Por ejemplo, en un caso de usuarios múltiples, en ocasiones explotando la disposición de conversión a través del mismo terminal, el registro/inicio de sesión puede tener lugar en el dispositivo 202 y en el servidor 208, posiblemente de manera automática en base a la información almacenada en el dispositivo 202 y la configuración actual. Además,

durante la etapa de puesta en marcha 302, la configuración del proceso de conversión se puede cargar o cambiar, y los valores de los parámetros que determinan, por ejemplo, las preferencias del usuario (algoritmos deseados de procesamiento de la voz, asociaciones entre la interfaz de usuario y los comandos de control, procedimiento de codificación, etc.) pueden establecerse. Aún más, el dispositivo 202 puede negociar con el servidor 208 los detalles de un escenario de cooperación preferible en el etapa 302, tal como se ha descrito anteriormente.

En el etapa 304, se inicia la captura de la señal de audio que incluye la voz a convertir, es decir, el transductor(es) del dispositivo 202 comienza(n) a traducir la vibración acústica de entrada en una señal eléctrica digitalizada con un convertidor A/D, que puede ser implementado como un chip separado o en combinación con el transductor(es). La señal primero será localmente capturada en el dispositivo 202 en su totalidad antes de ejecutar cualquier etapa adicional del procedimiento, o la captura se ejecuta simultáneamente con un número de etapas del procedimiento posteriores después de que se haya llevado a cabo primero la amortiguación mínima necesaria de la señal, por ejemplo. En 306 se muestra que el dispositivo 202 está configurado para monitorizar un comando de control comunicado al mismo, a través de los medios de control de entrada, simultáneamente con la captura de la señal de voz, en el que el comando de control determina uno o más elementos tales como signos de puntuación u otros elementos, opcionalmente simbólicos, y tareas de forma opcional. En el caso de recibir una orden de control, que se comprueba en 308, la naturaleza y el momento de la misma se verifican y se almacenan en 310, tal como se ha descrito anteriormente. La voz y los posibles comandos de control pueden ser controlados continuamente (téngase en cuenta la línea de trazos 315) hasta la recepción de un comando de parada, por ejemplo.

La etapa 312 se refiere al intercambio de información opcional con otras entidades, tal como el servidor 208. En un ejemplo, el dispositivo 202 registra la señal de audio y los posibles comandos de control, después de lo cual se transmiten al servidor 208 para la ejecución remota de al menos parte del proceso de conversión. En otra realización, el dispositivo 202 almacena de forma intermedia, y sustancialmente en tiempo real, transmite los datos de audio y de control al servidor 208. En ese escenario, el bloque 312 también podría colocarse en el grupo de bloques 304-310.

La etapa 314 se refiere a las tareas de realización de la conversión de voz a texto, en el que cada marca de puntuación u otro elemento determinado por el comando de control se coloca, por lo menos lógicamente, en una posición del texto correspondiente a la comunicación inmediata relativa a la señal de voz para cultivar la voz con el procedimiento de conversión de texto. El bloque 316 marca el final de la ejecución del procedimiento.

La figura 3b describe, a modo de ejemplo, un diagrama de flujo de un procedimiento de acuerdo con el ejemplo de la figura 2b. Los bloques 302, 304, y 316 son las etapas que corresponden en gran medida con las correspondientes de la figura 3a. En 318, si se aplica el reparto de tareas o la canalización de datos desde el dispositivo 202 hacia el servidor 208, los datos que representan la señal de voz capturada pueden ser transferidos en consecuencia. En 320, las tareas de conversión de voz texto se ejecutan, cuyo resultado posiblemente incluye una o más porciones con opciones de conversión múltiples, tal como se ha revisado anteriormente. En 322, al menos parte del resultado de la conversión que incluye las opciones para las una o más opciones puede ser transferido al dispositivo 202, en el caso de que la conversión fuera al menos parcialmente ejecutada en el servidor 208. En 324, una o más opciones son reproducidas por una sola porción, preferiblemente de forma audible, mediante el dispositivo 202 u otro dispositivo objetivo que recibe los datos de los resultados, y la respuesta del usuario se monitoriza 326. Los bloques 324, 326 pueden incorporar diversas opciones de repetición o reproducción, opcionalmente ajustables. Por ejemplo, el tono de reproducción (hombre, mujer, tono, volumen, etc.) y el tipo (reproducción del texto completo, incluyendo las porciones antes mencionadas, o partes más específicas, incluyendo las porciones, etc.), puede proporcionarse como opciones seleccionables. Tras la recepción de la selección del usuario 328 se incorpora, en 330, en el resultado de la conversión, que puede referirse a la supresión de las otras opciones y la adaptación de la selección como un texto estándar entre las palabras circundantes. Las etapas 324-330 se pueden repetir para el resto de las porciones con varias opciones de conversión, véase el número de referencia 331 que ilustra este procedimiento. Por ejemplo, la totalidad del texto puede ser reproducido inicialmente sustancialmente a partir de la selección anterior, o la reproducción puede comenzar desde la proximidad de la siguiente opción.

La figura 3c muestra un diagrama de flujo relativo a la edición de la señal y al intercambio de datos potencialmente teniendo lugar en el contexto de la presente invención. En esta etapa de escenario de ejemplo 304 también se puede indicar la codificación opcional de la señal de intercambio e información entre el dispositivo 202 y el servidor 208, si por lo menos parte de la señal se va a almacenar en el servidor 208 y el montaje se lleva a cabo de forma remota desde el dispositivo 202, o la edición se produce en piezas de datos que se transfieren entre el dispositivo 202 y el servidor 208. Como una alternativa al servidor 208, alguna otra entidad preferida podría ser utilizada como mero almacenamiento temporal de datos, si el dispositivo 202 no

contiene suficiente memoria para el propósito. Por lo tanto, aunque no se ilustra en la mayor medida por razones de claridad, las etapas presentadas en la figura 3c pueden comprender la transferencia de datos adicional entre el dispositivo 202 y el servidor 208/otra entidad, y la ruta explícitamente visualizada es simplemente una opción directa.

Las etapas 302, 304, y 316 en gran medida se ajustan a las etapas correspondientes de las figuras 3a y 3b, pero en el etapa 332 la señal se visualiza en la pantalla del dispositivo 202 para su edición. Las técnicas de visualización utilizadas pueden ser modificables por el usuario tal como revisa en la descripción de la figura 2c. El usuario puede editar la señal para cultivarla para que sea más relevante para el proceso de reconocimiento, e introducir las funciones de inspección de la señal preferidas (zoom/no zoom, diferentes representaciones paramétricas), funciones y algoritmos de configuración de la señal, e incluso volver a grabar/insertar/eliminar completamente las porciones necesarias. Cuando el dispositivo recibe una orden de edición por parte del usuario, ver el número de referencia 334, la acción asociada se lleva a cabo en la etapa de procesamiento 338, preferiblemente incluyendo también la funcionalidad "deshacer". Cuando el usuario se contenta con el resultado de la edición, el ciclo de etapas 332, 334, y 338 se deja atrás, y la ejecución del procedimiento continúa desde la etapa 336, que indica el intercambio de información entre el dispositivo 202 y el servidor 208. La información se refiere al proceso de conversión, e incluye, por ejemplo, la voz editada (opcionalmente también codificada).

Adicional o alternativamente (si por ejemplo, el dispositivo 202 o el servidor 208 es incapaz de hacerse cargo de una tarea), la señalización necesaria acerca de los detalles de reparto de tareas (negociaciones adicionales y parámetros relacionados, etc.) es transferida durante esta etapa. En la etapa 340, las tareas del proceso de reconocimiento se llevan a cabo según lo determinado por el escenario de negociación seleccionado. El número 344 se refiere al intercambio de información opcional para la transferencia de resultados intermedios, tales como voz procesada, parámetros de reconocimiento de voz calculados, porciones de texto o señalización adicional entre las entidades 202 y 208. Las porciones de texto separadas, posiblemente resultantes de la división de los trabajos se combinarán cuando esté listo construir el texto completo mediante el dispositivo 202, el servidor 208, o alguna otra entidad. El texto puede ser revisado por el usuario del dispositivo 202 y porciones del mismo ser objeto de correcciones, o incluso porciones de la voz original que corresponden al texto defectuoso producido pueden dirigirse a continuación para las rondas de conversión adicionales con una configuración opcionalmente modificada, si el usuario cree que vale la pena intentarlo. El texto final se puede considerar que se transfiere a la localización prevista (receptor, archivo, servicio adicional, etc.) durante la última etapa visualizada 316 que denota también el final de la ejecución del procedimiento. En el caso de la salida (texto traducido, voz sintetizada, etc.) desde el servicio adicional se transmite hacia adelante, la entidad de servicio adicional se basa en abordar el mensaje de orden de servicio recibido desde parte emisora, por ejemplo, el dispositivo 202 o el servidor 208, o bien devolver la salida de vuelta para que sea entregada en adelante a otra ubicación.

Un diagrama de señalización de la figura 4 describe una opción para la transferencia de información opcional entre el dispositivo 202 y el servidor 404. Cabe señalar, sin embargo, que las señales presentadas reflejan sólo un caso algo básico en el que múltiples rondas de conversión, etc. no se utilizan. La fecha 402 corresponde a la señal de audio, incluyendo la voz que se desea convertir. La señal 404 está asociada con una solicitud enviada al servidor 208 que indica el escenario de cooperación preferido para el proceso de conversión de voz a texto desde el punto de vista del dispositivo 202. El servidor 208 responde 406 con un reconocimiento que incluye una confirmación de la hipótesis aceptada, que puede diferir de la solicitada, determinada en base a los niveles de usuario, por ejemplo, y los recursos disponibles. El dispositivos 202 transmite los datos de los parámetros de reconocimiento de la voz, o al menos la porción de la señal de voz al servidor 208, tal como se muestra mediante la flecha 408. El servidor 208 realiza la parte negociada de la transformación y transmite los resultados al dispositivo 202 con los resultados posibles, incluyendo las opciones de conversión, o sólo reconoce su terminación 410. Los resultados pueden incluir opciones de conversión de resultados para determinadas porciones de texto. El dispositivo 202 transmite entonces el mensaje de aprobación/reconocimiento 412, opcionalmente, incluyendo todo el resultado de la conversión a procesar y/o transmitido a su destino final. El servidor 208 realiza opcionalmente al menos parte del procesamiento adicional y transmite la salida hacia adelante 414.

Un ejemplo no limitativo de un proceso de reconocimiento de voz que incluye un número de etapas se revisa a continuación para proporcionar a una persona experta una visión de la utilización de un aspecto, por ejemplo, de reparto de tareas de la invención actual. La figura 5 describe las tareas ejecutadas por un motor de reconocimiento de voz básico, por ejemplo, un módulo de software, en forma de un diagrama de flujo y bocetos ilustrativos relacionados con la función de tareas. Se hace hincapié en que la persona experta puede utilizar cualquier técnica de reconocimiento de voz adecuada en el contexto de la invención actual, y el ejemplo representado no se considerará como la única opción viable.

Las entradas del proceso de reconocimiento de voz de la señal de voz de forma digital (+ ruido adicional, si originalmente está presente y no se elimina durante la edición) que ya ha sido editada por el usuario del dispositivo 202. La señal se divide en marcos de tiempo con una duración de unas pocas decenas o centenares de milisegundos, por ejemplo, véase el número 502 y las líneas de puntos. La señal es analizada entonces en una base de fotograma a fotograma utilizando por ejemplo, análisis cepstral, durante el cual se calcula un número de coeficientes cepstrales mediante la determinación de una transformada de Fourier del marco y descorrelacionando el espectro con una transformada de coseno para recoger los coeficientes dominantes, por ejemplo, 10 primeros coeficientes por marco. También coeficientes derivados pueden ser determinados para la estimación de la expresión dinámica 504.

A continuación, el vector de la característica que comprende los coeficientes obtenidos y que representa el marco del habla se somete a un clasificador acústico, por ejemplo, un clasificador de red neuronal que se asocia con los vectores de características de diferentes fonemas 506, es decir, el vector de la característica está relacionado con cada fonema con una cierta probabilidad. El clasificador puede ser personalizado por los valores ajustables o los procedimientos de formación discutidos anteriormente.

En varios ejemplos de la presente invención, el clasificador, y el procedimiento de reconocimiento de voz en general, pueden formarse por separado para cada aplicación particular, basados en el vocabulario/diccionario, tales como médico, de negocios, o vocabularios jurídicos, por ejemplo, para mejorar el rendimiento de reconocimiento. El contexto de reconocimiento puede ser seleccionable/ajustable, por ejemplo, por el usuario a través de parámetros de aplicación, tal como un parámetro cuyo valor se adapta al reconocedor del escenario correspondiente. Alternativamente, el proceso de reconocimiento puede ser el mismo en cada escenario de uso independientemente del contexto.

En varios ejemplos de la presente invención, el procedimiento de reconocimiento también puede estar adaptado a cada idioma fuente, de tal manera que el usuario puede seleccionar el idioma aplicado, por ejemplo, a través de un interruptor de software que está funcionalmente acoplado con el reconocedor interior, por ejemplo. La selección del idioma puede alterar las reglas mediante las cuales el reconocedor analiza la voz de entrada de acuerdo con las características específicas de cada idioma, tales como definiciones de fonemas.

A continuación, las secuencias de fonemas se pueden construir mediante la concatenación de los fonemas, que subyacen a los vectores de características, posiblemente pueden ser analizados posteriormente con un HMM (Model de Markov oculto) u otro descodificador adecuado que determina la trayectoria 508 del fonema más probable (y el correspondiente elemento de nivel superior, palabra, por ejemplo) (formación de una frase: "esto parece ..." en la figura) a partir de las secuencias, por ejemplo mediante la utilización de un modelo de lenguaje dependiente del contexto léxico y/o gramatical y el vocabulario relacionado. Esta trayectoria se llama a menudo una trayectoria de Viterbi y se maximiza la probabilidad a posteriori para la secuencia en relación con el modelo probabilístico dado. El proceso de reconocimiento de voz puede incluir la determinación de múltiples opciones seleccionables por el usuario para ciertas porciones de texto, si las probabilidades asociadas no difieren considerablemente. Se obtienen comandos de control que definen la puntuación, por ejemplo, o las opciones de reconocimiento confirmadas por el usuario, pueden ser utilizados para la sección de la voz de entrada y el texto resultante, y, opcionalmente, para alterar las probabilidades de las opciones de reconocimiento circundantes. Mediante la aplicación de la puntuación obtenida, la selección y la información de contexto, por ejemplo, el proceso de reconocimiento de hecho puede proporcionar mejores resultados también como semántica del idioma, entrada adicional del usuario y/o sintaxis o gramática (en el sentido más general) se puede tomar en cuenta al determinar un resultado correcto del reconocimiento.

Ponderando sobre todo el aspecto de reparto de tareas, el intercambio podría tener lugar entre las etapas 502, 504, 506, 508 y/o incluso dentro de las mismas. En una opción, el dispositivo 202 y el servidor 208 pueden, en base a parámetros/reglas predeterminadas o negociaciones dinámicas/en tiempo real, asignar las tareas detrás de las etapas de reconocimiento 502, 504, 506, y 508, de tal manera que el dispositivo 202 se encarga de una serie de etapas (por ejemplo, 502) después de lo cual el servidor 208 ejecuta las etapas restantes (504, 506, y 508, respectivamente). Alternativamente, el dispositivo 202 y el servidor 208 ejecutarán todas las etapas, pero sólo en relación a una porción de la señal de voz, en cuyo caso las porciones convertidas de voz a texto se combinan finalmente mediante el dispositivo 202, el servidor 208, o alguna otra entidad para establecer el texto completo. Sin embargo, en una alternativa, las dos opciones anteriores pueden ser explotadas simultáneamente, por ejemplo, el dispositivo 202 se encarga de por lo menos una tarea para la señal de voz en conjunto (por ejemplo, la etapa 502) debido, por ejemplo, a un nivel de servicio actual a definir explícitamente, y que también ejecuta las etapas restantes para una pequeña porción de la voz concurrente con la ejecución de las mismas etapas restantes para el resto de la voz mediante el servidor 208. Esta división de tareas flexible puede originarse a partir de tiempo basado en la optimización de la expresión general para el proceso de conversión de texto, es decir, se estima que mediante la división aplicada, el dispositivo 202 y el servidor 208

finalizará sus tareas de forma sustancialmente simultánea y, por lo tanto, el tiempo de respuesta percibido por el usuario del dispositivo 202 se reduce al mínimo desde el lado de servicio. Sistemas de reconocimiento de voz modernos pueden llegar a un nivel de reconocimiento decente si la señal de voz de entrada es de buena calidad (libre de las perturbaciones y de ruido de fondo, etc.), pero el índice puede disminuir en condiciones más difíciles. Por lo tanto algún tipo de edición, comandos de control, y/u opciones que el usuario puede seleccionar tal como se han discutido pueden mejorar considerablemente el rendimiento del motor de reconocimiento de voz básica y la traducción de voz a texto en general.

La figura 6 muestra una opción para los componentes básicos del dispositivo electrónico 202, tal como un ordenador, un terminal móvil o una PDA, con capacidades de comunicaciones internas o externas. La memoria 604, dividida entre uno o más chips de memoria física, comprende el código necesario, por ejemplo, en forma de un programa/aplicación de ordenador 612 para permitir capturar la voz, almacenar, editar, o al menos la conversión parcial de voz a texto (~ motor de reconocimiento de voz), y otros datos 610, por ejemplo, la configuración actual, voz de forma digital (opcionalmente encriptada) y datos de reconocimiento de voz. La memoria 604 también se puede referir a una tarjeta de memoria extraíble, preferiblemente, un disquete, un CD-ROM o un medio de almacenamiento fijo, tal como un disco duro. La memoria 604 puede ser, por ejemplo, ROM o RAM por naturaleza. Medios de procesamiento 602, por ejemplo, una unidad de procesamiento/control tal como un microprocesador, un DSP, un microcontrolador o un chip lógico programable, que comprende opcionalmente una pluralidad de (sub-)límites de cooperación o paralelos se requieren para la ejecución real del código almaceno en la memoria 604. Una pantalla 606 y el teclado/almohadilla 608 u otros medios entrada de control aplicables (por ejemplo, una pantalla táctil o una entrada de control de voz) proporcionan al usuario del dispositivo 202 un control del dispositivo y medios de visualización de datos (~ interfaz de usuario). Los medios de entrada de voz 616 incluyen un sensor/transductor, por ejemplo, un micrófono y un convertidor A/D, para recibir una señal de entrada acústica y para transformar la señal acústica recibida en una señal digital. Los medios de transferencia de datos inalámbricos 614, por ejemplo, un transceptor de radio (GSM, UMTS, WLAN, Bluetooth, infrarrojos, etc.) son necesarios para la comunicación con otros dispositivos.

La figura 7 describe un diagrama de bloques correspondiente del servidor 208. El servidor comprende una unidad de control 702 y una memoria 704. La unidad de control 702 para controlar el motor de reconocimiento de voz y otras funcionalidades del servidor 208, incluyendo el intercambio de información de control, que en la práctica puede llevarse a cabo a través de los medios de entrada/salida de datos 714/718 u otros medios de comunicación, puede implementarse como una unidad de procesamiento o una pluralidad de unidades de cooperación como los medios de procesamiento 602 del dispositivo electrónico móvil 202. La memoria 704 comprende la aplicación del lado del servidor 712 para ser ejecutada por la unidad de control 702 para llevar a cabo al menos algunas de las tareas del un proceso general de conversión de voz a texto, por ejemplo un motor de reconocimiento de voz. Véase el párrafo anterior para ver ejemplos de posibles implementaciones de memoria. Aplicaciones y procesos opcionales 716 pueden proporcionarse para implementar servicios adicionales. Los datos 710 incluyen datos de voz, parámetros de reconocimiento de voz, configuraciones, etc. Por lo menos alguna información requerida puede estar localizada en una instalación de almacenamiento remoto, por ejemplo, una base de datos, a la que el servidor 808 tiene acceso a través de, por ejemplo, medios de entrada de datos 714 y medios de salida 718. Los medios de entrada de datos 714 comprenden, por ejemplo, una interfaz/adaptador de red (Ethernet, WLAN, Token Ring, ATM, etc.) para recibir los datos de voz e información de control enviada por el dispositivo 202. Del mismo modo, los medios de salida de datos 718 están incluidos, por ejemplo, para transmitir los resultados de la distribución de tareas hacia adelante. En la práctica, los medios de entrada de datos 714 y los medios de salida 718 pueden combinarse en una única interfaz multidireccional accesible mediante la unidad de control 702.

El dispositivo 202 y el servidor 208 se puede realizar como una combinación de software a medida y un hardware más genérico, o, alternativamente, a través de hardware especializado, tal como chips de lógica programable.

El código de aplicación, por ejemplo, la aplicación 612 y/o 712, que define un producto de programa de ordenador para la ejecución de la presente invención puede ser almacenado y entregado en un medio portador tal como un disquete, un CD, un disco duro o una tarjeta de memoria. El programa o software también puede ser entregado en una red de comunicaciones o un canal de comunicaciones.

El alcance de la invención se puede encontrar en las siguientes reivindicaciones. Sin embargo, los dispositivos utilizados, las etapas del procedimiento, el comando de control o detalles de opciones de conversión, etc. pueden depender de un caso de uso particular, convergiendo a las ideas básicas presentadas anteriormente, tal como aprecia un lector experto.

REIVINDICACIONES

1. Dispositivo electrónico para facilitar un procedimiento de conversión de voz a texto, que comprende:

- unos medios de entrada de voz para obtener una señal de voz digital,

- unos medios de entrada de control para comunicar un comando de control relativo a la señal de voz digital, mientras se obtiene la señal de voz digital,

- unos medios de procesamiento para asociar temporalmente el comando de control con un instante de tiempo sustancialmente correspondiente en la señal de voz digital a la que se dirige el comando de control,

en el que el comando de control determina uno o más signos de puntuación, símbolos u otros elementos de control que implican la manipulación del texto, para colocarse físicamente, tal como por ejemplo en el caso de dichos signos de puntuación y símbolos, o al menos lógicamente, a través de la manipulación del texto en el caso de dichos otros elementos de control, en una ubicación del texto correspondiente al instante de comunicación relativo a la señal de voz digital, para procurar el procedimiento de conversión de voz a texto a nivel local, en cuyo caso el dispositivo también comprende un motor de reconocimiento de voz para realizar tareas de conversión de voz a texto, o de forma remota, en cuyo caso el dispositivo electrónico también comprende unos medios de transferencia de datos para el envío de datos digitales que representan la señal de voz digital y el comando de control a una entidad remota para la conversión, o mediante un procedimiento de conversión compartida entre el dispositivo electrónico y la entidad remota, en cuyo caso el dispositivo electrónico también comprende al menos parte del motor de reconocimiento de voz y dichos medios de transferencia de datos.

2. Dispositivo electrónico según la reivindicación 1, en el que el comando de control también determina una o más acciones predeterminadas, tales como una pausa de grabación de longitud predeterminada, que se realiza en respuesta a la obtención del comando de control.

3. Dispositivo electrónico según cualquier reivindicación anterior, que además comprende un motor de reconocimiento de voz para realizar tareas de conversión de voz a texto, adaptado para aplicar la información proporcionada por el comando de control en la producción del resultado de la conversión.

4. Dispositivo electrónico según cualquier reivindicación anterior, en el que dichos medios de control de entrada comprenden una serie de elementos de entrada, cada uno asociado con al menos uno de dichos uno o más signos de puntuación, símbolos u otros elementos de control que implican la manipulación del texto.

5. Dispositivo electrónico según cualquier reivindicación anterior, que comprende un sintetizador de texto a voz y unos medios de salida de audio, y que se configura para, al obtener al menos un resultado parcial de la conversión de voz a texto que incluye una porción convertida, tal como una o más palabras o frases, que comprende múltiples, dos o más, opciones del resultado de la conversión seleccionables por el usuario, para reproducir, a través de dichos medios de salida de audio, una o más de dichas opciones para dicha porción, y para comunicar, a través de dichos medios de control de entrada, una selección de un usuario de dichas múltiples opciones seleccionables por el usuario para permitir la confirmación de un resultado de conversión deseado para dicha porción.

6. Servidor para realizar al menos una parte de la conversión de voz a texto, siendo el servidor operable en una red de comunicaciones, comprendiendo el servidor:

- unos medios de entrada de datos para recibir datos digitales enviados por un dispositivo de terminal, representando dichos datos digitales la señal de voz, y uno o más comandos de control, cada comando asociado temporalmente con un cierto instante de tiempo en los datos digitales y determinando uno o más signos de puntuación, símbolos u otros elementos de control que implican la manipulación del texto, y

- al menos parte de un motor de reconocimiento de voz para llevar a cabo las tareas de conversión de los datos digitales a texto, en el que el motor está adaptado para posicionar físicamente, como por ejemplo en el caso de dichos signos de puntuación y símbolos, o al menos lógicamente, a través de la manipulación del texto en el caso de dichos otros elementos de control, cada una de dichos signos de puntuación, símbolos u otros elementos de control impliquen la manipulación del texto en una ubicación del texto correspondiente al relación instante de tiempo determinado relativo a la señal de voz

representada por los datos digitales recibidos para cultivar el procedimiento de conversión de voz a texto al menos parcialmente adquiridos por el servidor.

7. Servidor según la reivindicación 6, que también comprende unos medios de salida de datos para transmitir por lo menos parte de la salida de las tareas realizadas a una entidad externa.

8. Servidor según cualquiera de las reivindicaciones 6-7, en el que dicha al menos parte de un motor de reconocimiento de voz está configurada para producir un resultado de conversión de voz a texto que incluye una porción convertida, tal como una o más palabras o frases, que comprende múltiples, dos o más , opciones del resultado de la conversión, cuando la corrección del resultado de la conversión se considera como incierto para la porción de acuerdo con el criterio predeterminado, y unos medios de salida de datos para comunicar el resultado de la conversión y al menos una indicación de las opciones al terminal u otro dispositivo remoto y, opcionalmente, activar el terminal que comprende un sintetizador de texto a voz y unos medios de salida de audio, u otro dispositivo remoto, para reproducir una forma audible una o más de dichas opciones para permitir la confirmación de un resultado de la conversión deseada para la porción por parte del usuario del terminal u otro dispositivo remoto en respuesta a la reproducción audible.

9. Procedimiento para la conversión de voz en texto, que comprende:

- obtener una señal de voz digital y un comando de control relativo a la misma de una manera de superposición temporal, en el que el comando de control determina uno o más signos de puntuación, símbolos u otros elementos de control que implican la manipulación del texto,

- asociar el comando de control con un instante de tiempo que corresponde sustancialmente en la señal de voz digital a la que se dirige el comando de control, y

- realizar una conversión de voz a texto, en la que cada signo de puntuación, símbolo u otro elemento de control implican la manipulación del texto determinado por el comando de control está físicamente posicionado, como por ejemplo en el caso de dicho signos de puntuación y símbolos, o al menos lógicamente, a través de la manipulación del texto en el caso de dichos otros elementos de control, en una ubicación del texto correspondiente al instante de comunicación relativo a la señal de voz para procurar el procedimiento de conversión de voz a texto.

10. Procedimiento según la reivindicación 9, que además comprende:

- obtener un resultado de la conversión de voz a texto que incluye una porción convertida, tal como una o más palabras o frases, que comprende múltiples, dos o más, opciones del resultado de la conversión,

- reproducir audiblemente una o más de dichas opciones,

- obtener una confirmación del usuario de una de dichas una o más opciones, y

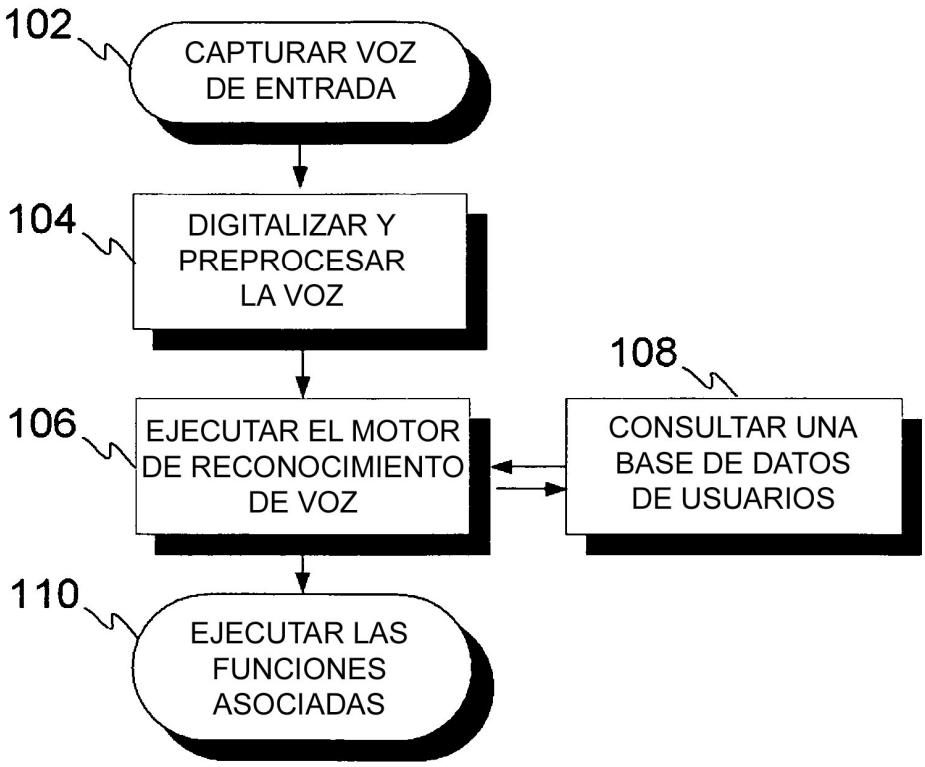
- seleccionar la conversión respecto a la porción convertida de acuerdo con la confirmación obtenida.

11. Programa ejecutable en un ordenador que comprende medios de código adaptados, cuando se ejecutan en un ordenador, para realizar las acciones del procedimiento tal como se definen en la reivindicación 9 ó 10.

12. Medio portador que comprende el programa ejecutable en un ordenador según la reivindicación 11.

13. Dispositivo electrónico según la reivindicación 1, que comprende un terminal móvil, una máquina de dictado, o un asistente digital personal (PDA).

14. Dispositivo electrónico o servidor según la reivindicación 1 ó 6, que además está configurado para, en respuesta a una entrada del usuario recibida, recibir una nueva voz o texto correspondiente y asociar dicha nueva voz o dicho texto correspondiente con los datos de voz o texto existentes convertidos a partir de los mismos, respectivamente, de manera que el resultado de la conversión obtenida comprende dicho texto correspondiente situado de acuerdo con la entrada del usuario.



TÉCNICA ANTERIOR

Figura 1

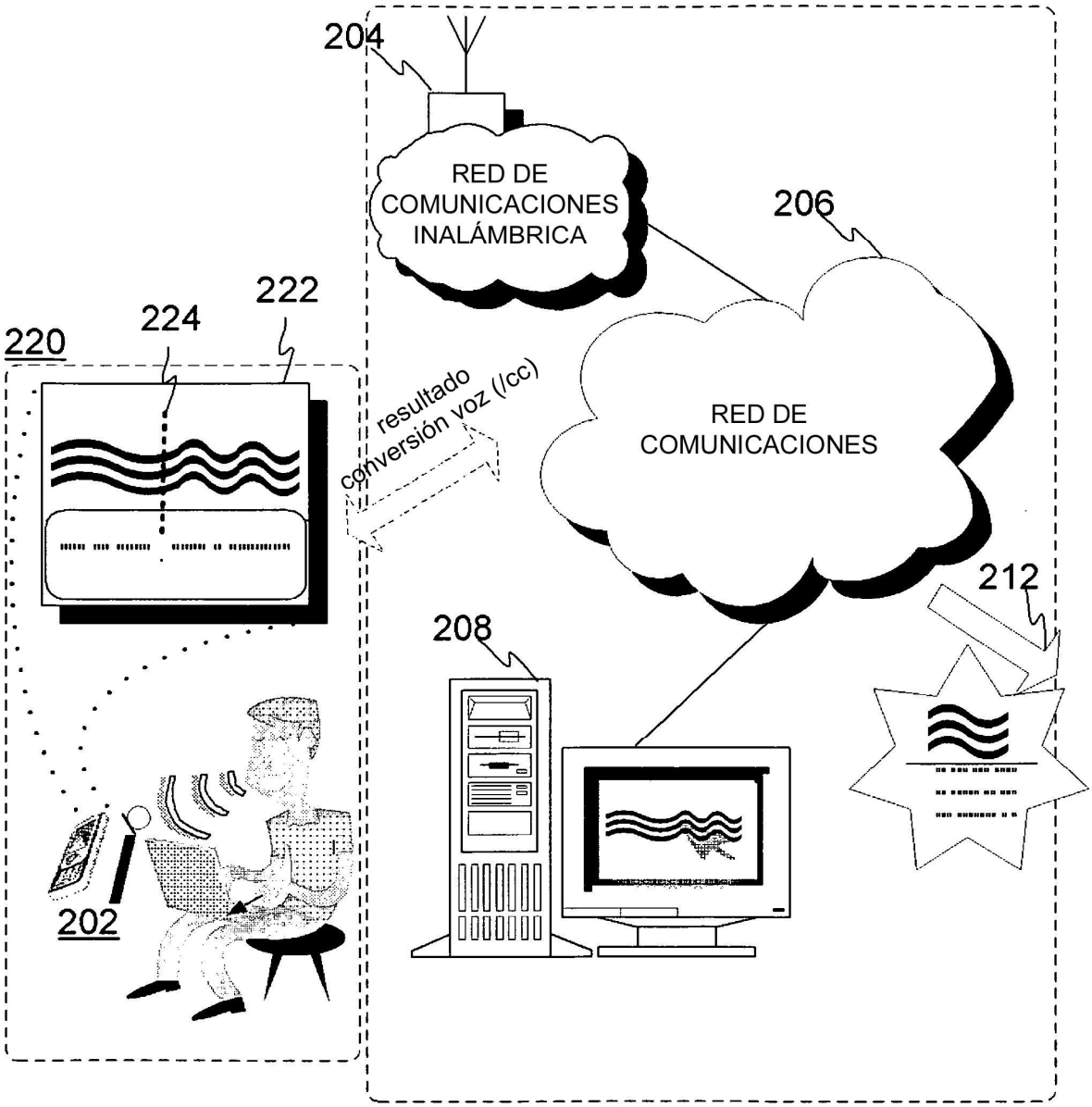

204
RED DE
COMUNICACIONES
INALÁMBRICA
206
222
224
220
resultado
conversión voz (/cc)
RED DE
COMUNICACIONES
212
208
202

Figura 2a

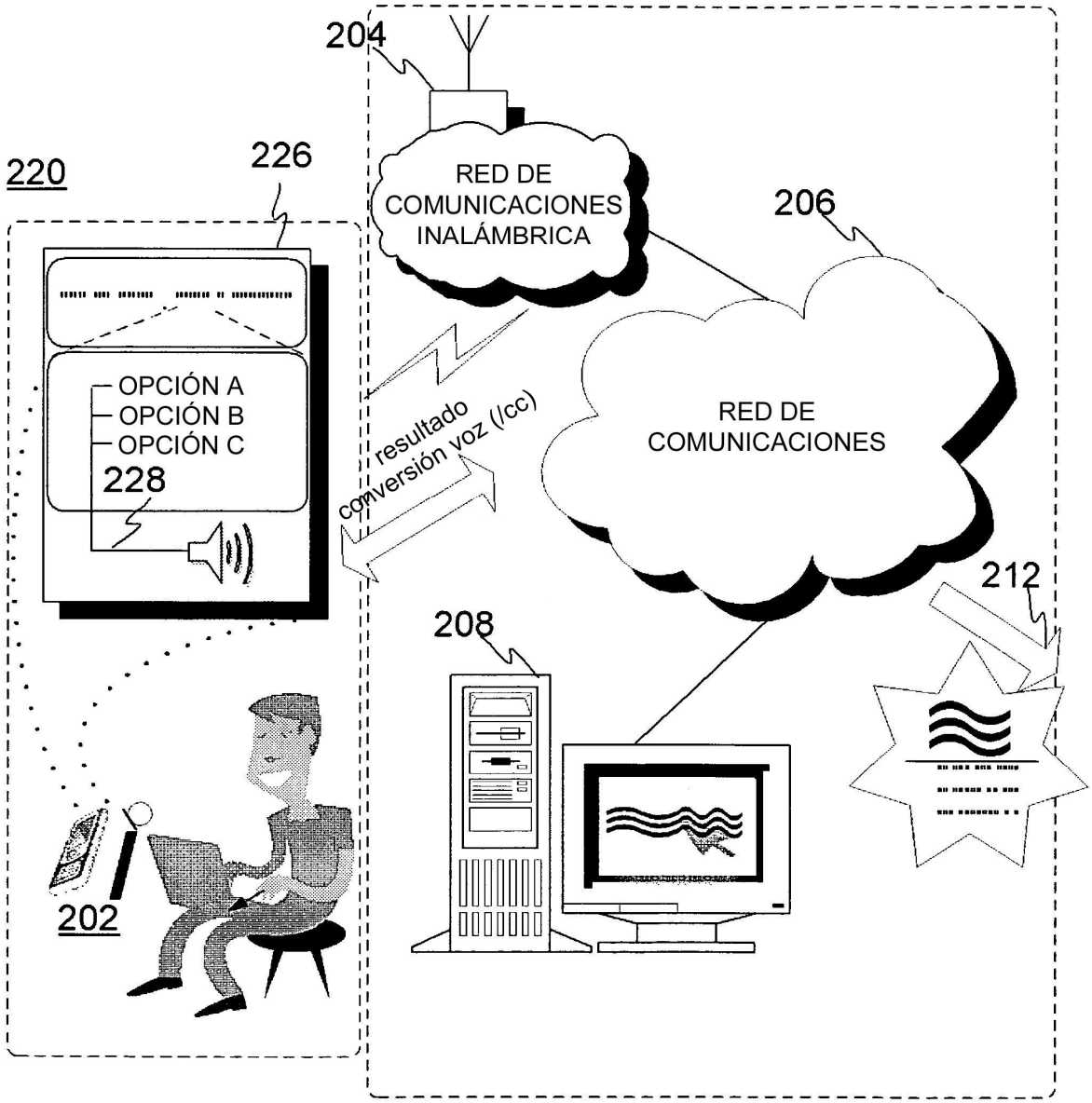

204
220
226
RED DE
COMUNICACIONES
INALÁMBRICA
206
OPCIÓN A
OPCIÓN B
OPCIÓN C
228
resultado
conversión voz (/cc)
RED DE
COMUNICACIONES
212
208
202

Figura 2b

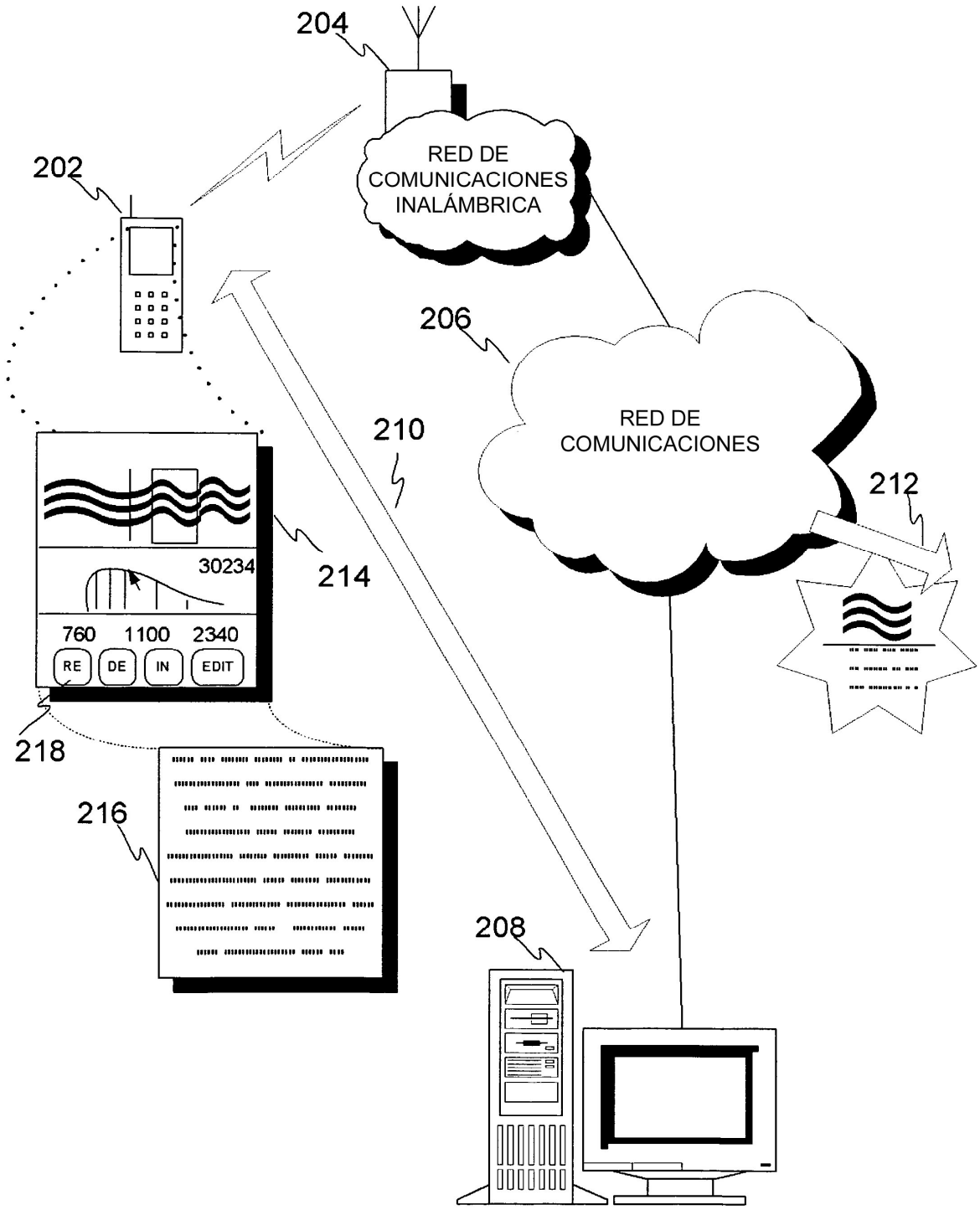

204
202
RED DE
COMUNICACIONES
INALÁMBRICA
206
RED DE
COMUNICACIONES
210
212
214
30234
760 1100 2340
RE DE IN EDIT
218
216
208

Figura 2c

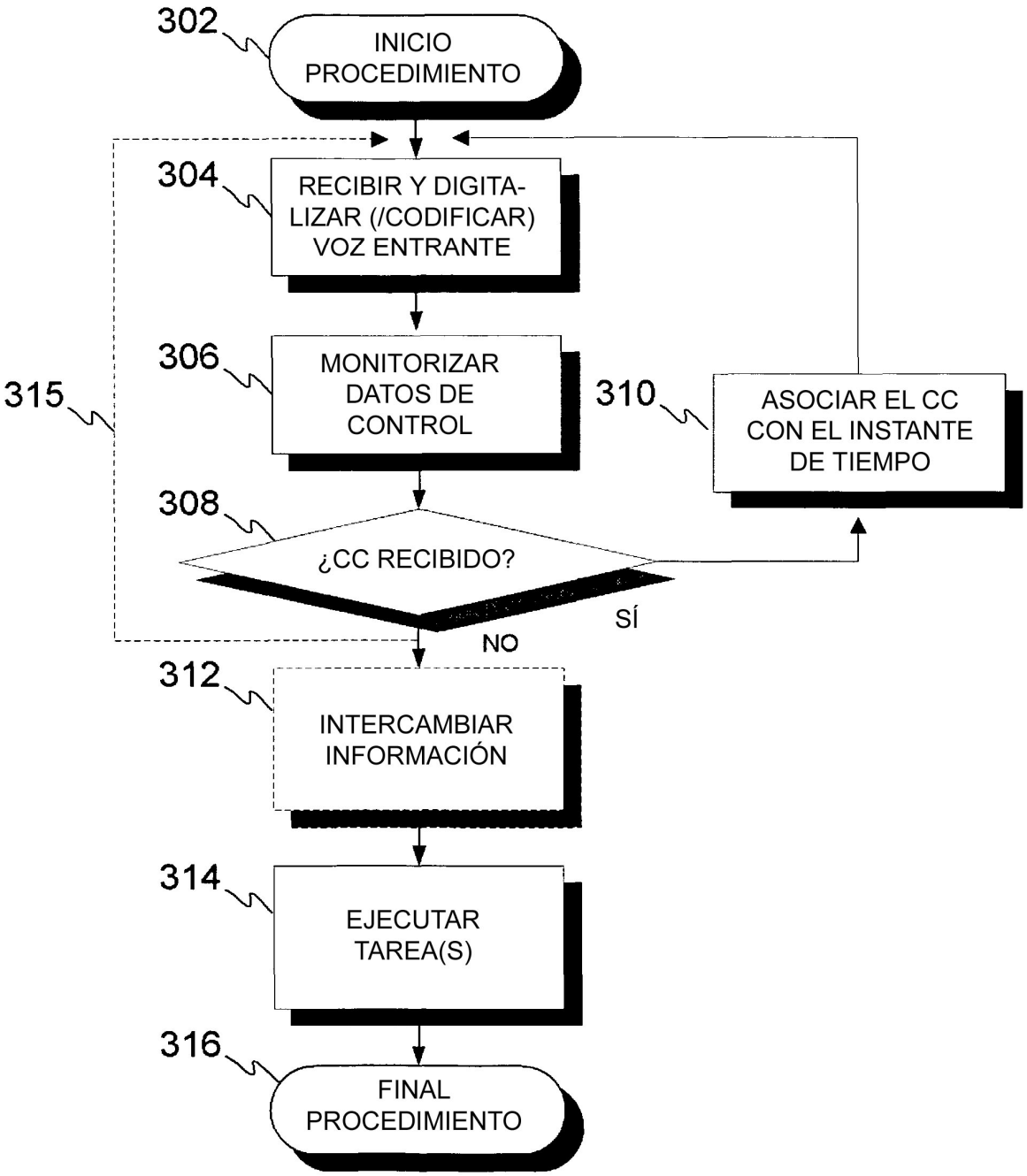

302
INICIO PROCEDIMIENTO
304
RECIBIR Y DIGITALIZAR (/CODIFICAR) VOZ ENTRANTE
306
MONITORIZAR DATOS DE CONTROL
315
310
ASOCIAR EL CC CON EL INSTANTE DE TIEMPO
308
¿CC RECIBIDO?
SÍ
NO
312
INTERCAMBIAR INFORMACIÓN
314
EJECUTAR TAREA(S)
316
FINAL PROCEDIMIENTO

Figura 3a

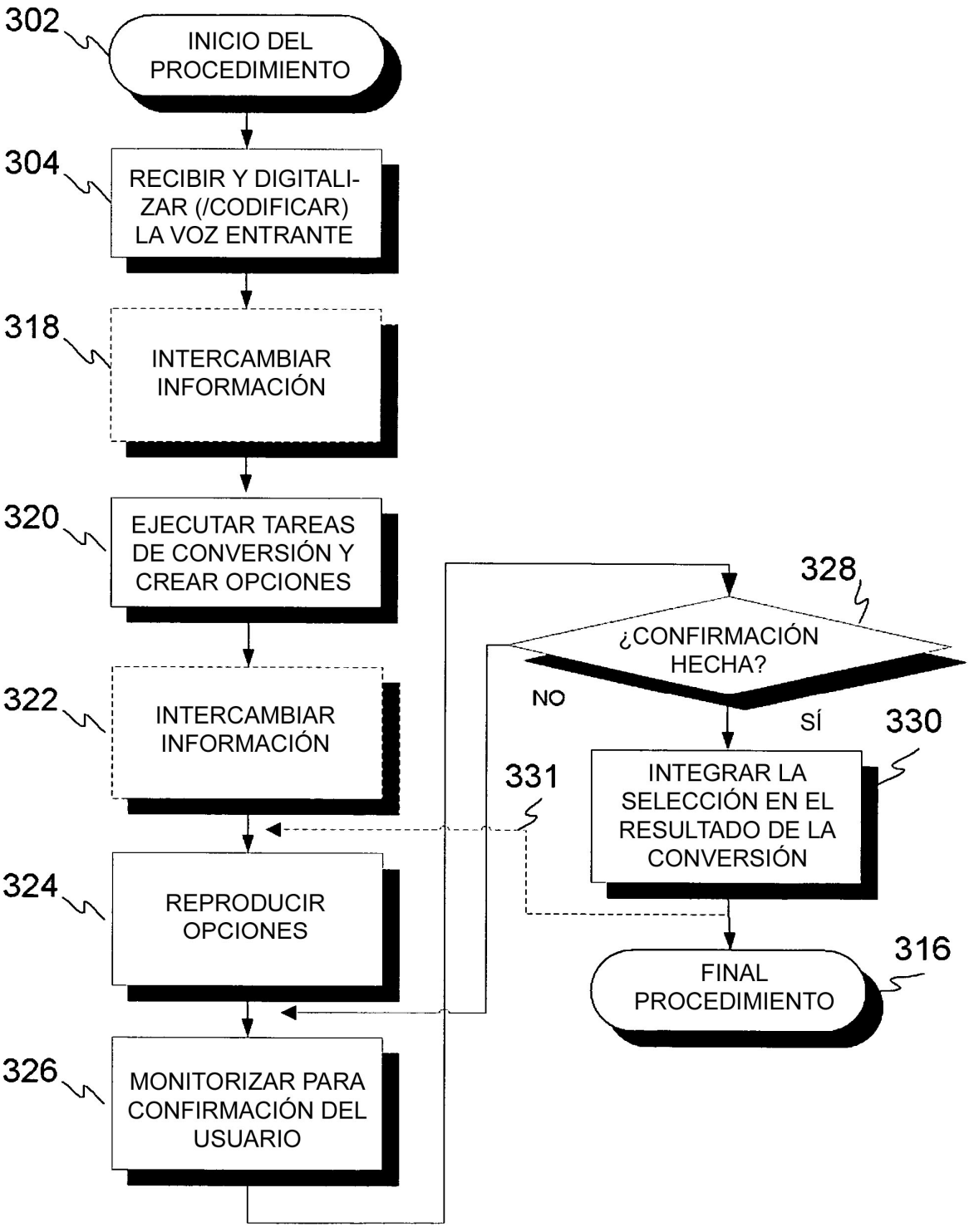

302
INICIO DEL PROCEDIMIENTO
304
RECIBIR Y DIGITALI-ZAR (/CODIFICAR) LA VOZ ENTRANTE
318
INTERCAMBIAR INFORMACIÓN
320
EJECUTAR TAREAS DE CONVERSIÓN Y CREAR OPCIONES
322
INTERCAMBIAR INFORMACIÓN
324
REPRODUCIR OPCIONES
326
MONITORIZAR PARA CONFIRMACIÓN DEL USUARIO
328
¿CONFIRMACIÓN HECHA?
NO
SÍ
330
331
INTEGRAR LA SELECCIÓN EN EL RESULTADO DE LA CONVERSIÓN
316
FINAL PROCEDIMIENTO

Figura 3b

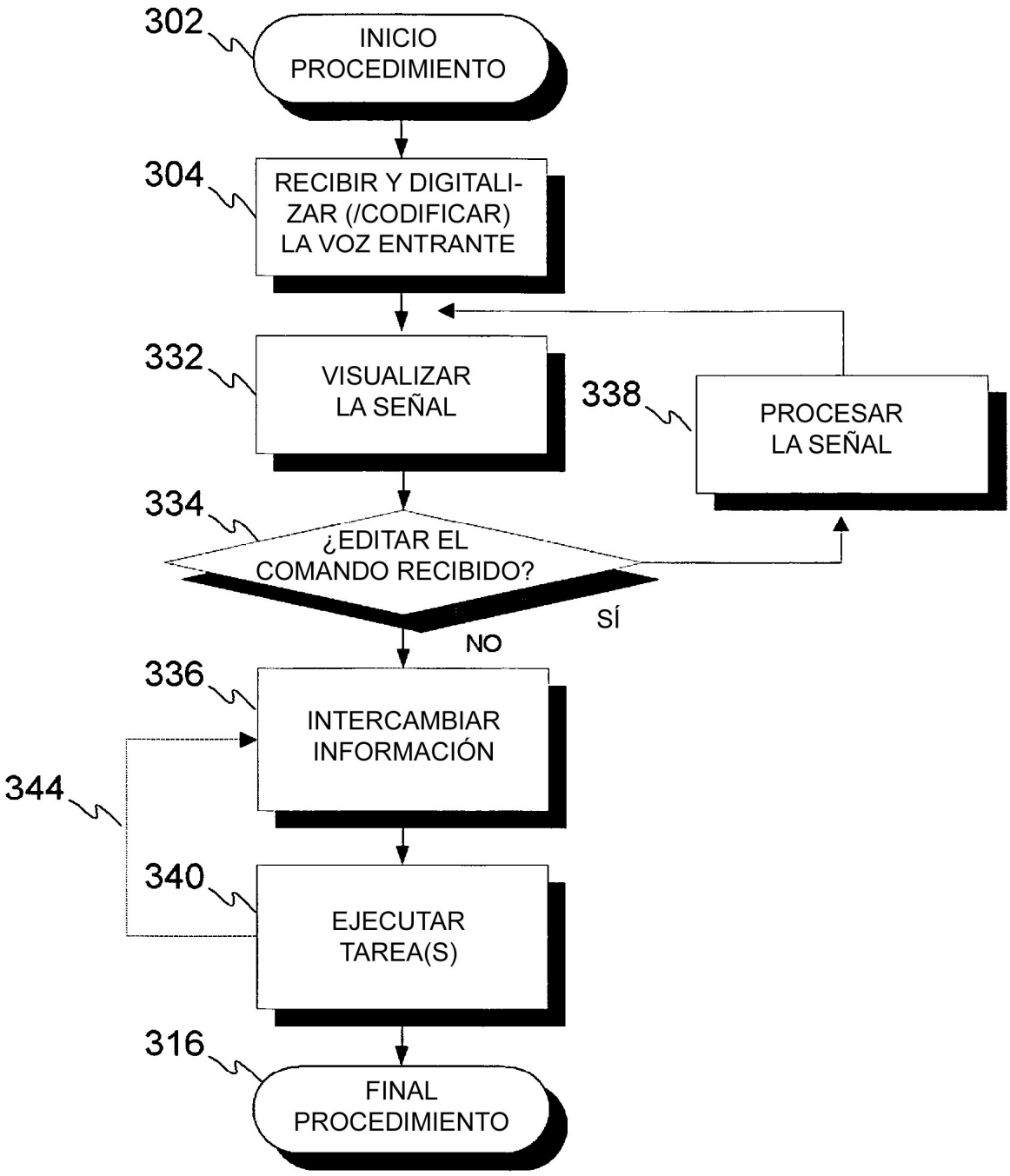

302
INICIO PROCEDIMIENTO
304
RECIBIR Y DIGITALI-ZAR (/CODIFICAR) LA VOZ ENTRANTE
332
VISUALIZAR LA SEÑAL
338
PROCESAR LA SEÑAL
334
¿EDITAR EL COMANDO RECIBIDO?
SÍ
NO
336
INTERCAMBIAR INFORMACIÓN
344
340
EJECUTAR TAREA(S)
316
FINAL PROCEDIMIENTO

Figura 3c

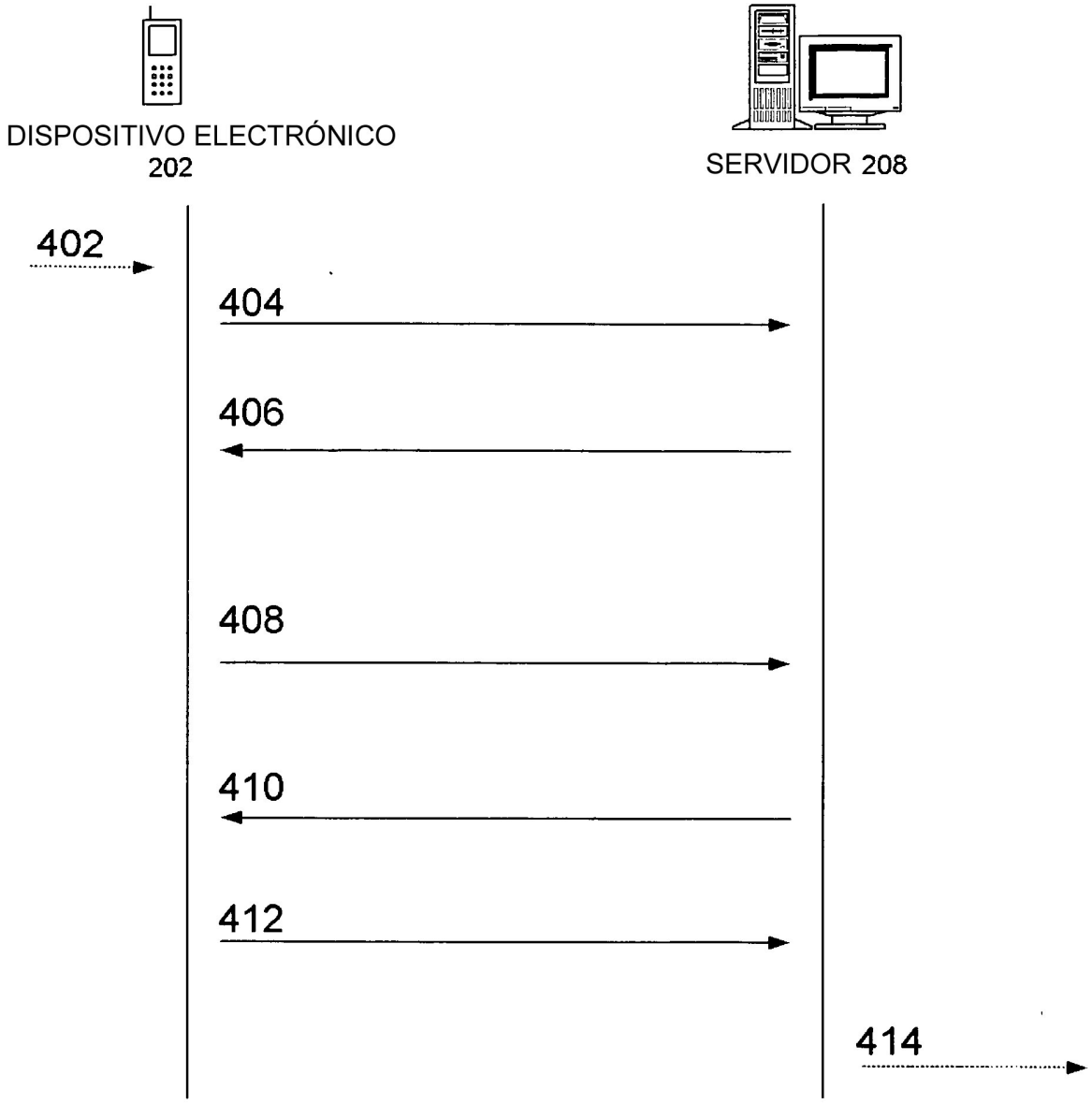

DISPOSITIVO ELECTRÓNICO
202
SERVIDOR 208
402
404
406
408
410
412
414

Figura 4

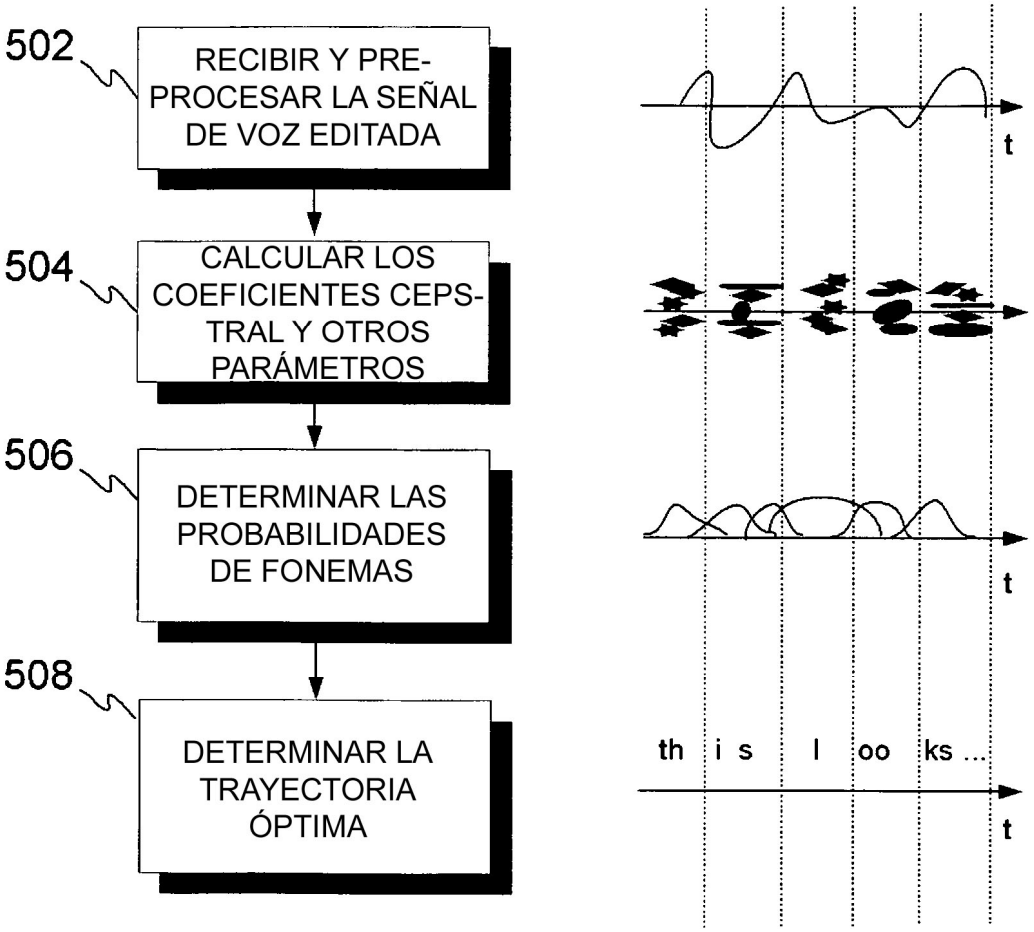

502
RECIBIR Y PRE-PROCESAR LA SEÑAL DE VOZ EDITADA
504
CALCULAR LOS COEFICIENTES CEPS-TRAL Y OTROS PARÁMETROS
506
DETERMINAR LAS PROBABILIDADES DE FONEMAS
508
DETERMINAR LA TRAYECTORIA ÓPTIMA
t
t
th i s l oo ks ...
t

Figura 5

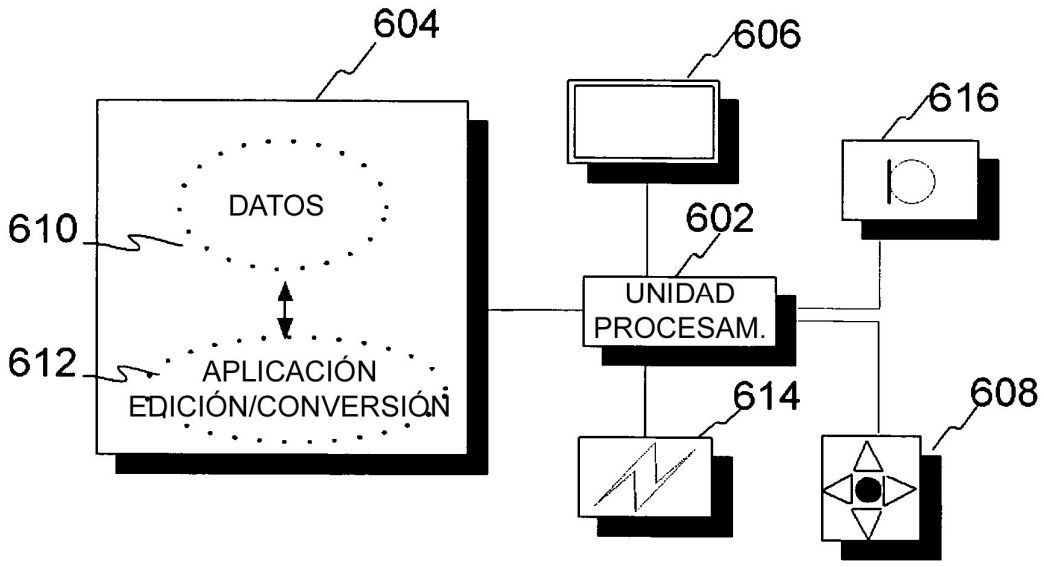


Figura 6

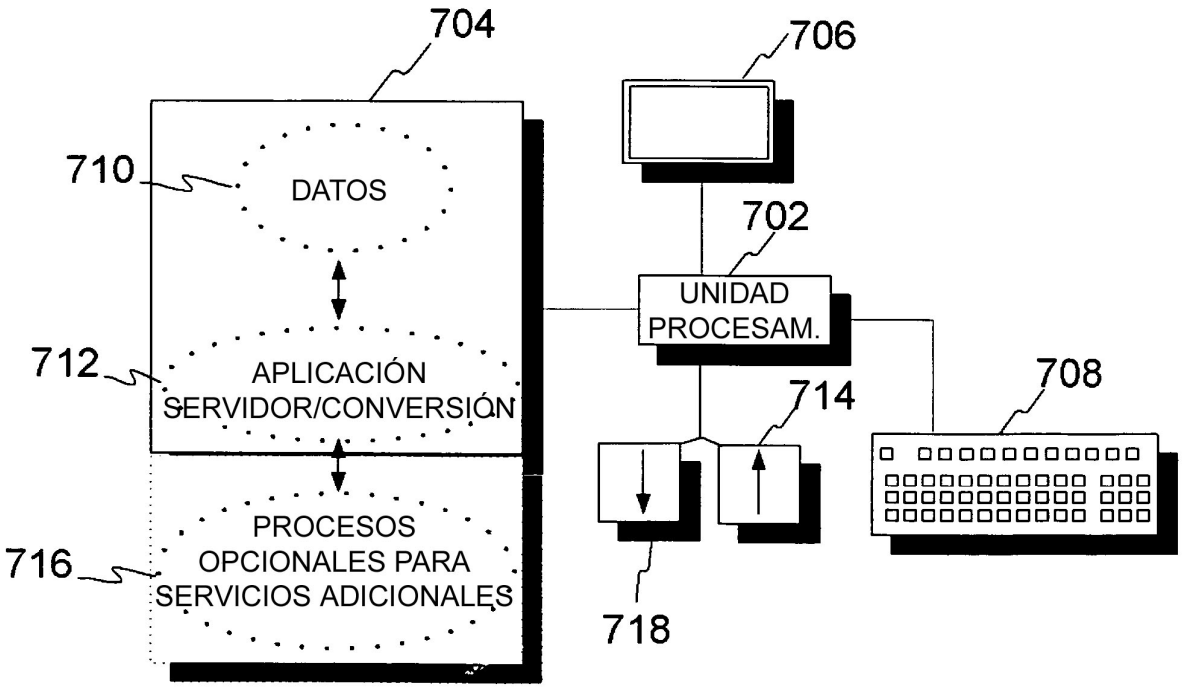


Figura 7