(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 359 430**

(51) Int. Cl.:
G10L 15/26 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **06743504 .0**
(96) Fecha de presentación : **27.04.2006**
(97) Número de publicación de la solicitud: **2036079**
(97) Fecha de publicación de la solicitud: **18.03.2009**

(54) Título: **Procedimiento, sistema y dispositivo para la conversión de la voz.**

(45) Fecha de publicación de la mención BOPI:
23.05.2011

(45) Fecha de la publicación del folleto de la patente:
23.05.2011

(73) Titular/es: **MOBITER DICTA Oy**
Museokatu 8 A 6
00100 Helsinki, FI

(72) Inventor/es: **Kurki-Suonio, Risto**

(74) Agente: **Carpintero López, Mario**

DESCRIPCIÓN

CAMPO DE LA INVENCIÓN

La presente invención versa en general acerca de dispositivos móviles y redes de comunicaciones. En particular, la invención se ocupa de una conversión interactiva de voz a texto y de servicios de traducción de idioma.

ANTECEDENTES DE LA INVENCIÓN

La actual tendencia en los terminales portátiles, por ejemplo de mano, conduce la evolución totalmente hacia interfaces de usuario intuitivas y naturales. Además de texto, en un terminal pueden grabarse imágenes y sonido (por ejemplo, voz), ya sea para la transmisión o para controlar una funcionalidad preferida local o remota (es decir, basada en la red). Además, puede transferirse información de carga útil por las redes celulares y fijas adyacentes como Internet como datos binarios que representan el texto, el sonido, las imágenes y el vídeo subyacentes. Así, los artilugios modernos en miniatura, como los terminales móviles o las PDA (agendas electrónicas) pueden tener medios versátiles de entrada de control, como un teclado o una botonera, un micrófono, diferentes sensores de movimiento o presión, etc., para proporcionar a los usuarios de los mismos una IU (interfaz de usuario) verdaderamente capaz de soportar el almacenamiento de datos muy diversificado y mecanismos de comunicaciones.

A pesar de que el salto en curso en la tecnología de las comunicaciones y la información, también algunas soluciones más tradicionales de almacenamiento de datos, como los dictáfonos, parecen mantener un considerable valor de facilidad de empleo, en especial en campos especializados como la ley y las ciencias médicas en los que se crean documentos regularmente en base, por ejemplo, a discusiones y encuentros verbales. Es probable que la comunicación verbal siga siendo el procedimiento más rápido y más cómodo de expresión para la mayoría de la gente, y dictar una nota en vez de mecanografiarla puede lograr ahorros de tiempo considerables. Este asunto también tiene un aspecto dependiente del idioma; por ejemplo, escribir en chino o japonés consume más tiempo, obviamente, que escribir en la mayoría de los idiomas occidentales. Además, los dictáfonos y los homólogos modernos de los mismos, como sofisticados terminales móviles y PDA con opción de grabación del sonido, pueden utilizarse hábilmente en conjunción con otras tareas, por ejemplo mientras se tiene una reunión o se conduce un coche, mientras que el mecanografiado manual normalmente requiere una parte importante de la atención de la persona que la realiza y, desde luego, no puede llevarse a cabo si se está conduciendo un coche, etc.

Hasta hace pocos años, los aparatos de dictado no han servido tan bien todas las necesidades públicas; sin duda, la información puede ser almacenada fácilmente incluso en tiempo real con solo grabar la señal de voz por medio de un micrófono, pero, a menudo, la forma final de archivo es textual y se encarga a alguien, por ejemplo a una secretaria, que manualmente limpie y convierta la señal de sonido grabado sin tratar en un registro final en un medio diferente. Desgraciadamente, tal acomodo requiere mucho trabajo adicional de conversión (consumo de tiempo). Otro problema importante asociado con los dictáfonos surge de sus antecedentes analógicos y de su IU simplista; modificar voz ya almacenada es engorroso y, dado que muchos dispositivos siguen utilizando cinta magnética como medio de almacenamiento, no pueden realizarse ciertas operaciones de edición, como insertar una porción de voz completamente nueva dentro de la señal almacenada originalmente. Por otra parte, los dictáfonos modernos que utilizan chips/tarjetas de memoria pueden incorporar opciones limitadas de edición de voz, pero la posible utilización sigue estando disponible únicamente a través de una IU más bien de difícil manejo que comprende únicamente una pantalla de LCD (pantalla de cristal líquido) de tamaño y calidad mínimos, etc. Transferir los datos almacenados de voz a otro dispositivo requiere a menudo un trasteo manual; es decir, el medio de almacenamiento (casete/tarjeta de memoria) debe ser movido físicamente.

Los sistemas informatizados de reconocimiento de voz llevan ya bastante tiempo a la disposición de una persona versada en la técnica. Típicamente, estos sistemas están implementados como características internas específicas a aplicaciones (integradas en un tratamiento de texto; por ejemplo, Microsoft Word, versión XP), aplicaciones dedicadas o módulos de aplicaciones conectables a un ordenador ordinario de sobremesa. El proceso de reconocimiento de voz implica varias etapas que están básicamente presentes en todos los algoritmos existentes; para una ilustración, véase la Figura 1. Concretamente, la fuente de voz emitida por una persona que habla es capturada 102, en primer lugar, por medio de un micrófono o un correspondiente transductor y convertida a forma digital con un necesario preprocesamiento 104 que puede referirse, por ejemplo, a un procesamiento de la dinámica. A continuación, la señal digitalizada es introducida en un motor 106 de reconocimiento de voz que divide la señal en elementos menores, como fonemas, en base a sofisticados procedimientos de extracción y análisis de características. El soporte lógico de reconocimiento también puede adaptarse 108 a cada usuario; es decir, las configuraciones del soporte lógico son específicas al usuario. Por últimos, los elementos reconocidos que forman la salida del motor de reconocimiento de voz, por ejemplo la información de control y/o el texto, son usados como entrada 110 para otros fines; puede simplemente mostrarse en la pantalla, almacenarse en una base de datos, traducirse a otro idioma, usarse para ejecutar una funcionalidad predeterminada, etc.

La publicación US6266642 da a conocer una unidad portátil dispuesta para llevar a cabo una traducción de lenguaje hablado para facilitar la comunicación entre dos entidades que no tienen ninguna lengua común. O bien el propio dispositivo contiene todos los soportes físicos y lógicos para ejecutar todo el proceso de traducción o meramente actúa

como interfaz remota que canaliza inicialmente la voz de entrada, utilizando una llamada ya sea de teléfono o de videoconferencia, a la unidad de traducción para su procesamiento, y después recibe el resultado de la traducción para la síntesis local de voz. La solución también comprende una etapa de procesamiento durante la cual se minimizan los errores de reconocimiento creando varios reconocimientos candidatos o hipótesis de entre las cuales el usuario puede seleccionar, por medio de una IU, la correcta o simplemente confirmar la selección predefinida.

La publicación US2003/0182113 da a conocer una disposición en la que un terminal móvil actúa como una sección de entrada de un sistema de reconocimiento de voz que consiste en el terminal móvil y al menos una entidad remota, como un ordenador de sobremesa conectado a la red, un ordenador central, un servidor de *web* o una pluralidad de ordenadores interconectados.

La publicación EP 1215659 da a conocer un sistema de reconocimiento de voz para permitir una conversión de (mensajes de) voz a texto. Se proporciona un medio de reconocimiento preliminar, como una red neural, según se integra en un dispositivo móvil que proporciona un código que representa los resultados del reconocimiento preliminar a una entidad remota por medio de una conexión inalámbrica.

A pesar de los muchos avances en las disposiciones mencionadas anteriormente y de que otras de la técnica anterior sugieren la superación de dificultades encontradas en el reconocimiento de voz y/o en los procesos de traducción automática, algunos problemas siguen sin resolver, especialmente en relación con los dispositivos móviles. Los problemas asociados con los dictáfonos tradicionales se ya describieron en lo que antecede del presente documento. Además, los dispositivos móviles como los terminales móviles y las PDA son, a menudo, aparatos de un tamaño relativamente pequeño y ligeros que no pueden incluir una pantalla de gran tamaño, una IU versátil, una capacidad/memoria de proceso de primera ni un transceptor de la más alta velocidad disponible, que están por lo común presentes en muchos dispositivos mayores, como los ordenadores de sobremesa. Tales características, aunque no son absolutamente necesarias para realizar el propósito original de un dispositivo portátil, o sea, transferir voz o almacenar información de calendario u otra información personal, serían beneficiosas desde el punto de vista de las conversiones de voz a texto y de la traducción automática. Las conversiones de formato de datos y las traducciones son muy exigentes en cuanto a las necesidades de cálculo y consumen mucho espacio en memoria; estos factores también llevarán de forma inevitable a un mayor consumo de batería, causando un problema más general de facilidad de uso con dispositivos móviles que siempre debería trabajar de manera ventajosa de forma independiente de su ubicación. Además, las prestaciones existentes de los dispositivos móviles de transferencia de la información pueden ser insuficientes para transferir todos los datos necesarios a la estación remota o desde la misma desde el mismo comienzo, o la capacidad de transferencia, aunque sea adecuada en teoría, puede no estar disponible para ese fin a plena escala debido, por ejemplo, a otra transferencia de información en curso con mayor prioridad.

RESUMEN DE LA INVENCIÓN

El objetivo de la presente invención es aliviar los defectos anteriormente mencionados encontrados en las actuales disposiciones de archivo de la voz y de conversión de la voz a texto. El objetivo se logra mediante una solución en la que un dispositivo móvil electrónico, por ejemplo un terminal móvil como un teléfono GSM/ UMTS/CDMA o una PDA equipada con un adaptador inalámbrico de comunicaciones comprenden una IU que permite al usuario, preferentemente mediante visualización, pero también a través de otros medios, editar la señal de voz antes de que sea expuesta al reconocimiento de voz propiamente dicho y a procesos opcionales, por ejemplo de traducción. Además, en la solución de la invención, la comunicación entre el dispositivo móvil y una entidad externa, por ejemplo un servidor de red que reside en la red a la que tiene acceso el dispositivo móvil, desempeña un papel importante. El dispositivo móvil y la entidad externa pueden estar configurados para dividir la conversión de voz a texto y acciones adicionales en base a varios valores de parámetros definibles por el usuario de forma ventajosa relativos, entre otros factores posibles, a cargas de procesamiento/memoria locales/remotas, estado de la batería, existencia de otras tareas y prioridad de las mismas, ancho de banda disponible para la transmisión, aspectos relativos a los costes, tamaño/duración de la fuente de la señal de voz, etc. El dispositivo móvil y la entidad externa pueden incluso negociar un escenario de cooperación adecuada en tiempo real en base a sus situaciones actuales; es decir, la compartición de tareas es un proceso dinámico. Estos asuntos se exponen con más detalle más adelante. Así, el proceso de conversión en su conjunto es, ventajosamente, interactivo entre el usuario del dispositivo móvil, el propio dispositivo móvil y la entidad externa. Además, el proceso de reconocimiento de voz puede personalizarse en relación con cada usuario; es decir, el motor de reconocimiento puede ser configurado o entrenado por separado para adaptarse a las características de la voz de aquel.

En un aspecto de la invención, un dispositivo móvil operable en una red de comunicaciones inalámbricas comprende:

- un medio de entrada de voz para recibir voz y convertir la voz en una señal digital de voz representativa,
- un medio de entrada de control para comunicar una orden de edición relativa a la señal digital de voz,
- un medio de procesamiento para llevar a cabo una tarea de edición de la señal digital de voz en respuesta a la orden de edición recibida,

- al menos parte de un motor de reconocimiento de voz para llevar a cabo tareas de conversión a texto de la señal digital de voz, y

- un transceptor para intercambiar información relativa a la señal digital de voz y la conversión de la misma de voz a texto con una entidad externa conectada funcionalmente a dicha red de comunicaciones inalámbricas.

En la solución anterior, la orden de edición y la tarea resultante pueden estar relacionados, sin limitación, con una de las siguientes opciones: supresión 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 la señal de voz, cambio de la amplitud de la señal de voz, cambio en el contenido espectral de la señal de voz, regrabación de una porción de la señal de voz. Preferentemente, el dispositivo móvil incluye un medio de visualización para visualizar la señal digital de voz, para que las órdenes de edición puedan relacionarse con la porción o porciones visualizadas de la señal.

El motor de reconocimiento de voz comprende una estructura, por ejemplo lógica de análisis, en forma de un soporte físico y/o lógico hechos a medida que se requieren para ejecutar al menos parte del proceso global de conversión de voz a texto partiendo de la voz en forma digital. Un proceso de reconocimiento de voz se refiere generalmente a un análisis de una señal de audio (que comprende voz) en base al cual la señal puede ser dividida ulteriormente en porciones menores y a la clasificación de las porciones. Así, el reconocimiento de voz permite y forma (al menos) una parte importante del procedimiento global de conversión de voz a texto de la invención, aunque la salida del simple motor de reconocimiento de voz también podría ser algo distinto del texto que representa textualmente la voz hablada; por ejemplo, en las aplicaciones de control de voz, el motor de reconocimiento de la voz asocia la voz de entrada con varias órdenes predeterminadas que el dispositivo anfitrión está configurado para ejecutar. Típicamente, todo el proceso de conversión incluye una pluralidad de etapas y, así, el motor puede llevar a cabo únicamente parte de las etapas o, de forma alternativa, la señal de voz puede ser dividida en “partes”, es decir, bloques, que son convertidas por entidades separadas. Cómo se lleva a cabo la compartición de tareas se expone más adelante. En un escenario mínimo, el dispositivo móvil puede ocuparse únicamente del procesamiento de la voz digital, en cuyo caso el dispositivo externo ejecutará las etapas más exigentes de cálculo, por ejemplo la de fuerza bruta.

En correspondencia con lo anterior, el intercambio de información se refiere a la interacción (recepción y/o transmisión de información) entre el dispositivo móvil y la entidad externa para ejecutar el proceso de conversión y procesos subsiguientes opcionales. Por ejemplo, la señal de voz de entrada puede transferirse, o bien completamente o parcialmente, entre los al menos dos elementos mencionados anteriormente, para que sea compartida la carga global de la tarea y/o las tareas específicas sean gestionadas por cierto elemento, tal como se ha mencionado en el párrafo inmediatamente anterior. Además, pueden transferirse diversos mensajes de parámetros, estado, acuse de recibo y de control durante la etapa de intercambio de información. En la descripción detallada se describen ejemplos adicionales. También se exponen formatos de datos adecuados para contener voz o texto.

En otro aspecto, un servidor operable en una red de comunicaciones comprende:

- un medio de entrada de datos para recibir una señal digital de datos enviada por un dispositivo móvil, representando dicha señal digital de datos voz o, al menos, parte de la misma, y para recibir una orden de edición de la voz a través del dispositivo móvil,

- al menos parte de un motor de reconocimiento de voz para llevar a cabo tareas de conversión a texto de una señal digital de datos,

- una unidad de control para intercambiar información de control con el dispositivo móvil, llevando a cabo una tarea de edición de la señal digital de voz en respuesta a la orden de edición recibida, y para determinar, en base a la información de control, las tareas que deben llevarse a cabo en la señal digital de datos recibida por dicha al menos parte del motor de reconocimiento de voz, y

- un medio de salida de datos para comunicar al menos parte de la salida de las tareas llevadas a cabo a una entidad externa.

En un aspecto adicional de la invención, un sistema para convertir voz a texto comprende un dispositivo móvil operable en una red de comunicaciones inalámbricas y un servidor conectado funcionalmente a dicha red de comunicaciones inalámbricas, en el que:

- dicho dispositivo móvil está configurado para recibir voz y convertir la voz en una señal digital de voz representativa, para recibir una orden de edición relativa a la señal digital de voz, para procesar la señal digital de voz según la orden de edición, para intercambiar 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 requeridas para llevar a cabo una conversión a texto de la señal digital de voz, y

- dicho servidor está configurado para recibir información relativa a la señal digital de voz y la conversión de voz a texto de la misma, y para ejecutar, en base a la información intercambiada, la parte restante de las tareas requeridas para llevar a cabo una conversión a texto de la señal digital de voz.

El “servidor” se refiere a una entidad, por ejemplo un aparato electrónico como un ordenador, que coopera con el dispositivo móvil de la invención para llevar a cabo la conversión de voz a texto y posibles procesos adicionales. La entidad puede estar incluida en otro dispositivo, por ejemplo una pasarela o un encaminador, o puede ser un dispositivo completamente separado o una pluralidad de dispositivos que forman la entidad servidora agregada de la invención.

Además, en otro aspecto, un procedimiento para convertir voz en texto tiene las etapas de:

- recibir, en un dispositivo móvil operable en una red inalámbrica, una fuente de voz, y convertir la fuente de voz en una señal digital de voz representativa,
- recibir una orden de edición relativa a la señal digital de voz por parte del dispositivo móvil,
- procesar la señal digital de voz según la orden de edición,
- intercambiar información relativa a la señal digital de voz y la conversión de voz a texto de la misma, y
- ejecutar, en base a la información intercambiada, al menos parte de las tareas requeridas para llevar a cabo una conversión de voz a texto de la señal digital de voz.

Preferentemente, la señal de voz digitalizada es visualizada en una pantalla del dispositivo móvil para que la edición pueda basarse también en la visualización.

La utilidad de la invención se debe a varios factores. En primer lugar, se pueden generar mensajes de forma textual con facilidad con fines de archivo y/o de comunicaciones hablando al dispositivo móvil de cada cual y, opcionalmente, editando la señal de voz por medio de la IU mientras el dispositivo y la entidad conectada de forma remota se ocupan automáticamente 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 medios diferentes (llamadas de voz, mensajes de texto, protocolos móviles de transferencia de datos, etc.), y la selección de un procedimiento actual de intercambio de información puede realizarse incluso dinámicamente en base, por ejemplo, a condiciones de la red. El texto y/o la voz editada resultantes pueden ser remitidos a un destinatario predeterminado utilizando una pluralidad de tecnologías y técnicas de comunicación diferentes, incluyendo Internet y redes móviles, redes informáticas de acceso restringido, correo de voz (síntesis de voz requerida para el texto resultado), correo electrónico, mensajes SMS/MMS, etc. El texto propiamente dicho puede proporcionarse en forma editable o de solo lectura. Los formatos aplicables de texto incluyen, por ejemplo, el ASCII sencillo (y otros juegos de caracteres), el formato Ms Word y el formato Adobe Acrobat.

El dispositivo móvil de la invención es, de forma ventajosa, un dispositivo o, al menos, está incorporado en un dispositivo que el usuario lleva consigo en cualquier ocasión y, por ello, no se introduce una carga adicional. Dado que el texto puede ser sometido ulteriormente a un motor de traducción automática, la invención también facilita la comunicación políglota. La editabilidad manual proporcionada de la señal de voz permite que el usuario verifique y cultive la señal de voz antes de la ejecución y de acciones ulteriores, lo que puede ahorrar al sistema un procesamiento innecesario y, ocasionalmente, mejorar la calidad de la conversión, ya que el usuario puede reconocer, por ejemplo, porciones inarticuladas en la señal de voz grabada y sustituirlas con versiones apropiadas. La compartición de tareas entre el dispositivo móvil y la entidad externa puede ser configurable y/o dinámica, lo que aumenta muchísimo la flexibilidad de la solución global, ya que pueden tenerse en cuenta los recursos disponibles de procesamiento/memoria para la transmisión de datos sin olvidar diversos aspectos adicionales, como el consumo de batería, el precio y los contratos del servicio, las preferencias del usuario, etc., incluso en tiempo real tras el aprovechamiento de la invención, tanto del dispositivo móvil como, específicamente, del usuario. El aspecto de la personalización de la parte del reconocimiento de voz de la invención aumenta, a su vez, la calidad de la conversión.

El núcleo de la presente invención puede ser expandido convenientemente por medio de servicios adicionales. Por ejemplo, pueden introducirse servicios manuales/automáticos de revisión ortográfica o de traducción de idioma/verificación de la traducción del texto, ya sea directamente por parte del operador del servidor o por terceros a los que el dispositivo móvil y/o el servidor transmiten los resultados de la conversión. Además, el lado servidor de la invención puede ser actualizado con los soportes físicos/lógicos más recientes (por ejemplo, soportes lógicos de reconocimiento) sin necesariamente suscitar una necesidad de actualizar el o los dispositivos móviles. En correspondencia con lo anterior, el soporte lógico móvil puede ser actualizado por medio de la comunicación entre el dispositivo móvil y el servidor. Desde un punto de vista del servicio, tal interacción abre nuevas posibilidades para definir una jerarquía exhaustiva de niveles de servicio. Dado que, típicamente, los dispositivos móviles, por ejemplo los terminales móviles, tienen capacidades diferentes y los usuarios de los mismos son capaces de gastar una cantidad variable de dinero (por ejemplo, en forma de costes de transferencia de datos o en tarifas directas por el servicio) para utilizar la invención, puede haber disponibles diversas versiones del soporte lógico móvil; la diferenciación puede implementarse por medio de bloqueo/activación de las características o como aplicaciones completamente separadas para cada nivel de servicio. Por ejemplo, en un nivel las entidades de la red se ocuparán de la mayor parte de las tareas de conversión, y el usuario está dispuesto a pagar por ello, mientras que en otro nivel el dispositivo móvil ejecutará una parte sustantiva del procesamiento, dado que incorpora las prestaciones necesarias y/o el usuario no quiere utilizar recursos externos para ahorrar costos o por alguna otra razón.

En una realización de la invención se presenta una disposición de la conversión de voz a texto siguiendo los principios explicados anteriormente. Una persona habituada a dictar notas utiliza su dispositivo móvil polivalente para capturar/editar una señal de voz y convertirla en texto en cooperación con un servidor de red. Se dan a conocer variaciones de la disposición básica.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

En lo que sigue se describe la invención en más detalle con referencia a los dibujos adjuntos, en los que:

la Fig. 1 ilustra un diagrama de flujo de un escenario típico de la técnica anterior, en el que un ordenador ordinario de sobremesa incluye un componente lógico de reconocimiento de voz;

la Fig. 2 visualiza una disposición de la conversión de voz a texto de la invención que comprende un dispositivo móvil y un servidor de red;

la Fig. 3 da a conocer un diagrama de flujo de una opción para llevar a cabo el procedimiento de la invención;

la Fig. 4 da a conocer un diagrama de señalización que muestra posibilidades de transferencia de información para implementar la presente invención;

la Fig. 5 representa interioridades del motor de reconocimiento de voz con un número de tareas;

la Fig. 6 es un diagrama de bloques de un dispositivo móvil de la invención;

la Fig. 7 es un diagrama de bloques de una entidad servidora de la invención.

DESCRIPCIÓN DETALLADA DE LA REALIZACIÓN DE LA INVENCIÓN

La Figura 1 ya fue objeto de reseña en conjunto con la descripción de la técnica anterior relacionada.

La Figura 2 da a conocer un esbozo de un sistema, únicamente a título de ejemplo, adaptación para llevar a cabo la disposición de conversión de la invención, tal como ha sido descrita en lo que antecede, bajo el control de un usuario que prefiere grabar sus mensajes y sus conversaciones en vez de mecanografiarlas en su dispositivo electrónico móvil polivalente, que proporciona una IU al resto del sistema. El dispositivo electrónico móvil 202, en lo sucesivo dispositivo 202, como un terminal móvil o una PDA con un medio de comunicaciones interno o externo, por ejemplo un transceptor de radiofrecuencia, es operable en una red 204 de comunicaciones inalámbricas, como una red celular o una red WLAN (LAN inalámbrica) capaz de intercambiar información con el dispositivo 202. Típicamente, las redes inalámbricas comprenden transceptores de radio denominados, por ejemplo, estaciones base o puntos de acceso para interconectarse con los dispositivos móviles. La comunicación inalámbrica puede referirse también al intercambio de otros tipos de señales aparte de meras señales de radiofrecuencia, incluyendo dichos tipos adicionales de señales, por ejemplo, las señales infrarrojas. Como término, la operabilidad se refiere en el presente documento a la capacidad de transferir información.

La red 204 de comunicaciones inalámbricas está conectada, además, a otras redes, por ejemplo una red 206 de comunicaciones (por cable), por medio de medios apropiados de interconexión, por ejemplo encaminadores o conmutadores. Conceptualmente, la red inalámbrica 204 puede estar también ubicada directamente en la red 206 de comunicaciones si no consiste más que en una interfaz inalámbrica para comunicarse con los terminales inalámbricos dentro del alcance. Internet es un ejemplo de red 206 de comunicaciones que también abarca una pluralidad de subredes.

Además, una entidad remota denominada servidor 208 reside en la red 206 de comunicaciones, o, al menos, está conectada a la misma a través de redes intermedias. El dispositivo 202 y el servidor 208 intercambian información 210 por medio de las redes 204, 206 para llevar a cabo el proceso global de conversión de voz a texto de la invención. El motor de reconocimiento de voz está situado en el servidor 208 y, opcionalmente al menos, en el dispositivo 202. El texto y/o la voz editada resultantes pueden ser entonces comunicados 212 a un destinatario remoto dentro o fuera de dichas redes de comunicaciones inalámbricas 204 y de comunicaciones 206, a un archivo electrónico (en cualquier red o dentro del dispositivo 202, por ejemplo en una tarjeta de memoria) o a una entidad de servicio que se ocupa del procesamiento ulterior, por ejemplo de la traducción de los mismos. De forma alternativa o adicional, el procesamiento ulterior puede llevarse a cabo en el servidor 208.

Los bloques 214, 216 representan capturas de pantalla potenciales del dispositivo 202 realizadas tras la ejecución del procedimiento global de conversión de voz. La captura 214 ilustra una opción para presentar visualmente, 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, en efecto, ser visualizada para su repaso o su edición sacando provecho de varios enfoques diferentes: la representación del dominio temporal de la señal puede dibujarse como una envolvente (véase la curva superior en la captura) o como un gráfico más tosco (por ejemplo, del tipo de voz activada/desactivada u otra segmentación del dominio temporal de resolución reducida, en cuyo caso la resolución reducida puede obtenerse

de la señal original, por ejemplo, dividiendo el intervalo de los valores originales de la misma en un número menor de subintervalos con valores de umbral limitados) en base a la amplitud o los valores de magnitud de la misma, y/o puede calcularse de la misma un espectro de potencia u otra parametrización de dominio frecuencial o alternativo (véase la curva inferior en la captura).

Incluso pueden aplicarse simultáneamente varias técnicas de visualización, con lo que, por ejemplo, mediante una funcionalidad de acercamiento (/alejamiento) o de otro tipo, pueden mostrarse en otro lugar de la pantalla cierta parte de la señal correspondiente a un intervalo temporal definido por el usuario o un subintervalo de valores preferidos de parámetros (véase las curvas superior e inferior de la captura 214 presentadas simultáneamente) con una resolución aumentada (/disminuida) o mediante una técnica alternativa de representación. Además de la o las representaciones de la señal, la captura 214 muestra diversos valores numéricos determinados durante el análisis de la señal, marcadores (rectángulo) y puntero (flecha, línea vertical) a la señal (porción) y funciones corrientes de edición o visualización de datos aplicadas o disponibles (véase el número de referencia 218). En caso de una pantalla táctil, el usuario, de forma ventajosa, puede pintar con el dedo o un estilete una zona preferida de la porción de señal visualizada (de forma ventajosa, la señal puede ser desplazada por el usuario si no llena la pantalla con una resolución preferida) y/o, presionando otra zona predeterminada, especificar una función que deba ser ejecutada en relación con la porción de la señal subyacente a la zona preferida. Puede proporcionarse al usuario una funcionalidad similar por medio de un medio de control más convencional, por ejemplo un puntero que se mueva por la pantalla en respuesta a la señal de control del dispositivo de entrada creado por un controlador de puntero, un ratón, un botón de botonera/teclado, un controlador direccional, un receptor de órdenes de voz, etc.

A partir de la señal visualizada, el usuario del dispositivo 202 puede reconocer rápidamente, requiriéndose únicamente una experiencia mínima, los sonidos separables, como palabras, y las posibles señales aberrantes (ruidos de fondo, etc.) contenidos en la misma y editar ulteriormente la señal para cultivarla para el subsiguiente proceso de reconocimiento de voz. Por ejemplo, si se muestra una envolvente de la representación del dominio temporal de la señal de voz, las porciones de menor amplitud a lo largo del eje del tiempo corresponden, con mucha probabilidad, al silencio o el ruino de fondo, mientras que los sonidos de voz contienen más energía. En el dominio frecuencial, los picos dominantes son debidos, a su vez, a los componentes de la señal de voz propiamente dicha.

El usuario puede introducir y comunicar órdenes de edición de la señal al dispositivo 202 por medio de la IU del mismo. Preferentemente, las funciones de edición asociadas con las órdenes permitirán una inspección y una revisión exhaustivas de la señal original, dándose a conocer por ello, a continuación, algunos ejemplos útiles.

La porción de la señal definida por el usuario (por ejemplo, ya sea seleccionada con marcadores/punteros móviles o “pintada” en la IU como en la pantalla táctil, según se ha explicado en lo que antecede) será sustituible con otra porción, ya sea que esté ya almacenada o que sea grabada en tiempo real. Así mismo, una porción será suprimible para que las porciones adyacentes restantes puedan unirse entre sí o para que la porción borrada sea sustituida con unos datos predeterminados que representen, por ejemplo, silencio o ruido de fondo de bajo nivel. En los extremos de la señal capturada es innecesario tal procedimiento de unión. Al usuario puede asignársele una posibilidad de alterar, por ejemplo unificar, la amplitud (relativa al volumen/sonoridad) y el contenido espectral de la señal, lo que puede llevarse a cabo a través de diferentes medios de control de la ganancia, algoritmos de normalización, un ecualizador, un controlador del margen dinámico (incluyendo, por ejemplo, una puerta de ruido, un expansor, un compresor, un limitador), etc. Los algoritmos de reducción de ruido para limpiar la señal de voz degradada del jaleo de fondo son más complejos que el uso de puertas de ruido, pero ventajoso siempre que la señal acústica original haya sido producida en condiciones ruidosas. Preferentemente, el ruido de fondo será al menos pseudoestacionario para garantizar una precisión adecuada de modelado. Los algoritmos modelan el ruido de fondo espectralmente o por medio de un filtro (coeficientes) y restan de la señal microfónica la estimación del ruido modelado ya sea en el dominio temporal o en el espectral. En algunas soluciones, la estimación de ruido se actualiza únicamente cuando un detector de actividad vocal (DAV) aparte notifica que no hay voz en la porción actualmente analizada de la señal. Generalmente, la señal puede clasificarse como que incluye solo ruido, solo voz o ruido + voz.

La aplicación de conversión puede almacenar varias funciones y varios algoritmos diferentes de edición de la señal que son seleccionables por el usuario como tales, y al menos algunos de ellos pueden ser personalizados adicionalmente por el usuario, por ejemplo a través de varios parámetros ajustables.

En la aplicación se incluye, preferentemente, la funcionalidad de cancelación, también denominada funcionalidad de “deshacer”, que se trata, por ejemplo, de una opción de programa para volver al estado de la señal antes de la última operación, para permitir que el usuario experimente sin riesgos con los efectos de diferentes funcionalidades mientras busca una señal editada óptima.

Siempre que la edición ocurre, al menos en parte, simultáneamente con el reconocimiento de voz, puede visualizarse en la pantalla del dispositivo 202 incluso únicamente el texto resultante hasta el momento. Esto puede requerir una transferencia de información entre el servidor 208 y el dispositivo 202, si el servidor 208 ha participado en la conversión de la porción particular de voz de la que se ha originado el texto resultante hasta el momento. Si no, se materializa la captura 216 después de completar la conversión de voz a texto. De manera alternativa, nunca se muestra el texto como tal al usuario del dispositivo 202, ya que, por defecto, es remitida directamente al destino de archivo o a un destinatario remoto, preferentemente dependiendo de las configuraciones definidas por el usuario.

Una configuración puede determinar si el texto es mostrado automáticamente en la pantalla del dispositivo 202 para su revisión, de nuevo opcionalmente junto con la señal de voz original o la editada, es decir, la señal de voz se visualiza como se ha descrito en lo que antecede, mientras que las porciones resultantes de texto, como las palabras, se muestran encima o debajo de la voz alineadas en relación con las correspondientes porciones de voz. Los datos necesarios para el alineamiento se crean como producto secundario en el proceso de reconocimiento de voz durante el cual la señal de voz ya se analiza en porciones. El usuario puede entonces determinar si está satisfecho con el resultado de la conversión o decidir editar adicionalmente las porciones preferidas de voz (incluso regrabarlas) y someterlas a una nueva ronda de reconocimiento mientras que mantiene intactas las porciones restantes, si las hay. Lo cierto es que este tipo de conversión recursiva de voz a texto consume más tiempo y recursos que el tipo de enfoque básico y directo, del tipo “editar una vez y convertir”, pero permite el logro de resultados más precisos. De manera alternativa, al menos parte del texto resultante puede ser corregida introduciendo manualmente correcciones para omitir rondas adicionales de conversión sin auténtica certidumbre de resultados más precisos.

Aunque la señal de audio introducida que comprende la voz es capturada en origen por el dispositivo 202 por medio de un sensor o un transductor como un micrófono y luego es digitalizada por medio de un convertidor A/D para la transmisión y/o el almacenamiento de la forma digital, incluso la fase de edición puede comprender una transferencia de información entre el dispositivo 202 y otras entidades como el servidor 208, tal como se anticipa con el enfoque recursivo anterior. A su vez, la señal digital de voz puede ser de un tamaño tan grande que no pueda ser almacenada de manera sensata en el dispositivo 202 como tal; por lo tanto, tiene que ser comprimida localmente, opcionalmente en tiempo real durante la captura, utilizando un codificador de audio dedicado a la voz o más genérico, como GSM, TETRA, G.711, G.721, G.726, G.728, G.729 o diversos codificadores de la serie MPEG. Además, o de forma alternativa, la señal digital de voz puede, tras la captura, ser transmitida directamente (incluyendo, no obstante, la necesaria acumulación intermedia) a una entidad externa, por ejemplo el servidor 208, para el almacenamiento y, opcionalmente, la codificación, y ser devuelta posteriormente al dispositivo 202 para su edición. En casos extremos, la edición tiene lugar en el servidor 208, de modo que el dispositivo 202 actúa principalmente como una interfaz remota para controlar la ejecución de las funciones de edición en el servidor 208 previamente explicadas. Para ese fin, tienen que transferirse entre las dos entidades 202, 208 tanto los datos de voz (para su visualización en el dispositivo 202) como la información de control (órdenes de edición).

El intercambio 210 de información en conjunto puede incorporar una pluralidad de características diferentes de la disposición de conversión. En un aspecto de la invención, el dispositivo 202 y el servidor 208 comparten las tareas relativas a la conversión de voz a texto. La compartición de tareas implica también inherentemente el intercambio 210 de información, dado que, al menos una porción de la voz (codificada opcionalmente) 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, al menos, tienen acceso a configuraciones para la compartición de tareas (por ejemplo, de funciones, algoritmos) con varios parámetros, que pueden ser definibles o fijados por el usuario (o, al menos, no libremente alterables por el usuario). Los parámetros pueden o bien determinar explícitamente cómo se dividen las tareas entre el dispositivo 202 y el servidor 208, o únicamente supervisar el proceso a través de varias reglas más genéricas que han de seguirse. Por ejemplo, ciertas tareas pueden ser llevadas a cabo siempre por el dispositivo 202 o por el servidor 208. Las reglas pueden especificar la compartición de la carga de procesamiento, en la que se determinan umbrales de carga o bien relativos o absolutos con adaptividad/lógica adicional opcional para las cargas tanto del dispositivo 202 como del servidor 208 para transferir generalmente parte del procesamiento y, así, de los datos fuente 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 un abono que incluye varios niveles de servicio, algunas características de conversión pueden estar deshabilitadas a un cierto nivel (inferior) de usuario bloqueándolas, por ejemplo, en la aplicación de conversión. La funcionalidad de bloqueo/desbloqueo puede llevarse a cabo a través de un conjunto de versiones diferentes del soporte lógico, de una característica de códigos de registro, de módulos adicionales descargables de soporte lógico, etc. En el caso de que el servidor 208 no pueda implementar algunas de las tareas permitidas de nivel inferior solicitadas por el dispositivo 202, por ejemplo durante una situación de sobrecarga del servidor o de interrupción del servicio en el servidor, puede enviar un mensaje de “falta de acuse de recibo” u omitir por completo el envío de cualquier respuesta (a menudo, tal como se presenta en la Figura 4, las faltas de acuse de recibo se envían realmente), para que el dispositivo 202 pueda deducir del acuse de recibo negativo o ausente que debe ejecutar las tareas por sí mismo siempre que sea posible.

El dispositivo 202 y el servidor 208 pueden negociar un escenario de cooperación para la compartición de tareas y el intercambio 210 de información resultante. Tales negociaciones pueden ser desencadenadas por el usuario (o sea, seleccionando una acción que lleve al comienzo de las negociaciones), de manera programada (una vez al día, etc.), tras el comienzo de cada conversión, o dinámicamente durante el proceso de conversión transmitiéndose información de parámetros entre sí, por ejemplo en conexión con un cambio en el valor de un parámetro. Los parámetros relativos a la compartición de tareas incluye información, por ejemplo, a uno o más de los siguientes: carga de procesamiento o de memoria actuales, estado de la batería o su capacidad máxima, el número de tareas distintas en ejecución (con prioridad mayor), el ancho de banda disponible para la transmisión, aspectos relacionados con el coste, como la tasa actual de transmisión de datos para la o las rutas de transferencia disponibles o el costo de uso del servidor por tamaño/duración de los datos de voz, el tamaño o la duración de la señal fuente de voz, los procedimientos disponibles de codificación/decodificación, etc.

El servidor 208 es, en la mayoría de los casos, superior al dispositivo 202 en cuanto a la potencia de procesamiento y la capacidad de memoria; por lo tanto, las comparaciones de carga serán relativas o a escala en todo caso. La lógica para llevar a cabo la compartición de tareas puede basarse en simples tablas de valores umbral que incluyan, por ejemplo, diferentes intervalos de valores de parámetros y las decisiones resultantes de la compartición de tareas. En la práctica, la negociación puede realizarse por medio del intercambio 210 de información para que o bien el dispositivo 202 o el servidor 208 transmitan información de estado al otro que determine una escenario optimizado de cooperación y devuelva el resultado del análisis para iniciar el proceso de conversión.

El intercambio 210 de información también cubre la transmisión del estado de la conversión (anuncios de ejecución lista/en curso de la tarea actual, aviso de interrupción del servicio, cifras de carga del servicio, etc.) y mensajes de señalización de acuse de recibo (recepción de datos realizada con/sin éxito, etc.) entre el dispositivo 202 y el servidor 208. Sin embargo, no siempre que se solucionan las asignaciones de compartición de tareas es obligatoria la transferencia de la señalización relacionada.

El intercambio 210 de información puede tener lugar con prácticas diferentes de comunicación, incluso múltiples simultáneamente (transferencia de datos en paralelo) para acelerar las cosas. En una realización, el dispositivo 202 establece una llamada de voz con el servidor 208 por la cual se transmite la señal de voz, o, al menos, parte de ella. La voz puede ser transferida en conexión con la fase de captura, o después de editarla por vez primera en el dispositivo 202. En otra realización, se usa un protocolo dedicado de transferencia de datos, como el GPRS, para la transferencia de voz y de otra información. La información puede ser encapsulada en diversos formatos de paquete/trama 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 combinarán en cualquiera de los dos dichos dispositivos 202, 208 para crear el texto final. Dependiendo de la naturaleza de la compartición (¿representan los resultados intermedios las porciones correspondientes del texto final?), los resultados intermedios puede ser transmitidos, de manera alternativa, como tales a una entidad receptora adicional, que puede llevar a cabo el proceso final de combinación aplicando información que le proporcionan para ese fin las entidades 202, 208.

Servicios adicionales, como revisión ortográfica, traducción automática/humana, verificación de la traducción o síntesis adicional de texto a voz, pueden localizarse en el servidor 208 o en otra entidad remota a la que se transmite el texto después de completar la conversión de voz a texto. En el caso de que los resultados intermedios anteriormente mencionados se refieran directamente a porciones del texto, las porciones pueden ser transmitidas de manera independiente inmediatamente después de su finalización, con la condición de que la respectiva información adicional para la combinación también se acabe transmitiendo.

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 ser personalizado para utilizar las características de voz individuales de cada usuario. Esto indica la introducción de las características en una base de datos local o remota accesible por el motor de reconocimiento en base, por ejemplo, a la ID del usuario; las características pueden ser obtenidas, de manera conveniente, entrenando al motor o bien proporcionándose al motor parejas libremente seleccionadas de muestra de voz/texto correspondiente o pronunciando las expresiones que el motor está configurado a solicitar de cada usuario en base a, por ejemplo, un compromiso predefinido (dependiente del idioma) entre maximizar la versatilidad y el valor representativo del espacio de la información y minimizar el tamaño de la misma. En base al análisis de los datos de entrenamiento, el motor determina a continuación las configuraciones personalizadas, por ejemplo los parámetros del reconocimiento, que han de usarse en el reconocimiento. Opcionalmente, el motor ha sido adaptado para actualizar continuamente la información del usuario (~perfiles de usuario) utilizando la información de retorno recogida; pueden analizarse las diferencias entre el texto final corregido por el usuario y al texto final producido automáticamente.

La Figura 3 da a conocer un diagrama de flujo de un procedimiento según la invención. En la etapa 302 de inicio del procedimiento se llevan a cabo las acciones iniciales que permiten la ejecución de las etapas ulteriores del procedimiento. Por ejemplo, se lanzan las aplicaciones necesarias, una o más, relativas al proceso global de conversión de voz a texto, incluyendo la edición, en el dispositivo 202 y se activa el respectivo servicio en el lado del servidor 208. En el supuesto caso de que el usuario del dispositivo 202 deseara una fase de reconocimiento personalizado, la etapa 302 incluye opcionalmente el registro o el inicio de sesión en el servicio. Esto también tiene lugar siempre que el servicio está destinado únicamente a usuarios registrados (servicio privado) y/u ofrece una pluralidad de niveles de servicio. Por ejemplo, en un caso de múltiples usuarios que saquen provecho ocasionalmente de la disposición de conversión por medio del mismo terminal, el registro o el inicio de sesión puede tener lugar tanto en el dispositivo 202 como en el servidor 208, posiblemente automáticamente, en base a la información almacenada en el dispositivo 202 y en las configuraciones actuales. Además, durante la etapa 302 de inicio pueden cambiarse las configuraciones del proceso de conversión y fijarse los valores de los parámetros que determinan, por ejemplo, diversas preferencias del usuario (algoritmos por defecto de procesamiento de voz, procedimiento de codificación, etc.). Además, tal como se ha descrito en lo que antecede, el dispositivo 202 puede negociar con el servidor 208 un escenario preferente de cooperación en la etapa 302.

En la etapa 304 se inicia la captura de la señal de audio, incluyendo la voz que ha de ser convertida, es decir, el o los transductores del dispositivo 202 comienzan a traducir la vibración acústica de entrada a una señal eléctrica digitalizada

con un convertidor A/D que puede ser implementado como un chip aparte o combinado con el o los transductores. O bien la señal será capturada en primer lugar localmente en el dispositivo 202 en su conjunto antes de que se ejecute ninguna etapa adicional del procedimiento, o la captura discurre simultáneamente con varias etapas subsiguientes del procedimiento después de que se haya llevado a cabo primer el necesario almacenamiento intermedio mínimo de la señal. La etapa 304 también indica la codificación opcional de la señal y el intercambio de información entre el dispositivo 202 y el servidor 208, si al menos parte de la señal ha de ser almacenada en el servidor 208 y la edición tiene lugar de manera remota con respecto al dispositivo 202, o si la edición se produce en segmentos de datos que son transferidos entre el dispositivo 202 y el servidor 208. Como alternativa al servidor 208, podría usarse alguna entidad distinta como un mero almacenamiento temporal de datos si el dispositivo 202 no contiene suficiente memoria para ese fin. Por lo tanto, aunque no se ilustren en la más amplia extensión por razones de claridad, las etapas 302-310 presentadas en la Figura 3 pueden comprender una transferencia adicional de datos entre el dispositivo 202 y el servidor 208 u otra entidad, y la trayectoria visualiza explícitamente es, simplemente, la opción más directa.

En la etapa 306 la señal se visualiza en la pantalla del dispositivo 202 para su edición. Las técnicas de visualización utilizadas pueden ser alterables por el usuario, como se reseñó en la descripción de la Figura 2. El usuario puede editar la señal para cultivarla para hacerla más relevante al proceso de reconocimiento e introducir funciones preferidas de inspección de la señal (acercamiento/alejamiento, diferentes representaciones paramétricas), funciones/algoritmos de formación de la señal, e incluso regrabar/insertar/borrar por completo las porciones necesarias. Cuando el dispositivo recibe del usuario una orden de edición (véase el número de referencia 308), la acción asociada se lleva a cabo en la etapa 310 del procesamiento, preferentemente incluyendo también la funcionalidad de “deshacer”. Cuando el usuario está satisfecho con el resultado de la edición, queda atrás el bucle de las etapas 306, 308 y 310, y la ejecución del procedimiento continúa a partir de la etapa 312, que indica el intercambio de información entre el dispositivo 202 y el servidor 208. La información se relaciones con el proceso de conversión e incluye, por ejemplo, la voz editada (opcionalmente, también codificada). Además, o de manera alternativa (si, por ejemplo, el dispositivo 202 o el servidor 208 son incapaces de ocuparse de una tarea), durante esta etapa se transfiere la señalización necesaria en cuanto a detalles de la compartición de tareas (negociación ulterior y parámetros relacionados, etc.). En la etapa 314 se llevan a cabo las tareas del proceso de reconocimiento según lo determinado por el escenario de negociación seleccionado. El número 318 se refiere al intercambio opcional de información adicional para transferir resultados intermedios como la voz procesada, los parámetros calculados de reconocimiento de voz, las porciones de texto o señalización adicional entre las entidades 202 y 208. Las porciones separadas de texto que posiblemente resulten de la compartición de tareas se combinarán cuando el dispositivo 202, el servidor 208 o alguna otra entidad estén listos para construir el texto completo. El texto puede ser revisado por el usuario del dispositivo 202 y porciones del mismo ser sometidas a correcciones, o incluso porciones de la voz original correspondientes al texto defectuoso producido pueden seleccionarse para rondas adicionales de conversión con configuraciones opcionalmente corregidas, si el usuario cree que merece la pena intentarlo. Puede considerarse que el texto final es transferido a la ubicación deseada (destinatario, archivo, servicio adicional, etc.) durante la última etapa 316 visualizada, lo que denota también el fin de la ejecución del procedimiento. En el caso de que se remita la salida (texto traducido, voz sintetizada, etc.) desde el servicio adicional, la entidad del servicio adicional la abordará en base al mensaje recibido de petición de servicio desde el remitente, por ejemplo el dispositivo 202 o el servidor 208, o les devolverá la salida para que sea entregada a otra ubicación.

El diagrama de señalización de la Figura 4 da a conocer una opción para la transferencia de información entre el dispositivo 202 y el servidor 404 en el espíritu de la invención. Sin embargo, debería hacerse notar que las señales presentadas reflejan solamente un caso un tanto básico en el que no se utilizan múltiples rondas de conversión, etc. La flecha 402 corresponde a la señal de audio que incluye la voz que ha de ser convertida. La señal 404 está asociada con una solicitud enviada al servidor 208 que indica el escenario preferida de cooperación para el proceso de conversión de voz a texto desde el punto de vista del dispositivo 202. El servidor 208 responde 406 con un acuse de recibo que incluye una confirmación del escenario aceptado, que puede diferir del solicitado, determinado en base, por ejemplo, a los niveles de usuario y a los recursos disponibles. El dispositivo 202 transmite datos o porciones de parámetros de reconocimiento de voz de la señal de voz al servidor 208, tal como se muestra mediante la flecha 408. El servidor 208 lleva a cabo la parte negociada del procesamiento y transmite los resultados al dispositivo 202 o simplemente da acuse de su terminación 410. Acto seguido, el dispositivo 202 transmite un mensaje 412 de aprobación/acuse de recibo que incluye opcionalmente todo el resultado de la conversión para que sea procesado adicionalmente y/o transmitido al destino final. Opcionalmente, el servidor 208 lleva a cabo al menos parte del procesamiento adicional y transmite la salida 414 a la etapa siguiente.

A continuación se presenta una vista previa de un ejemplo no limitante de un proceso de reconocimiento de voz que incluye varias etapas para proporcionar a una persona versada una comprensión en cuanto a la utilización del aspecto de compartición de tareas de la presente invención. La Figura 5 da a conocer tareas ejecutadas por un motor básico de reconocimiento de voz, por ejemplo un módulo de soporte lógico, en forma de un diagrama de flujo y bosquejos ilustrativos relativos a la función de las tareas. Se recalca que la persona versada puede utilizar cualquier técnica adecuada de reconocimiento de voz en el contexto de la presente invención, y no se considerará que el ejemplo representado sea la única opción viable.

El proceso de reconocimiento de voz introduce la señal de voz de formato digital (+ruido adicional, si estaba presente en origen y no se eliminó durante la edición) que ya ha sido editada por el usuario del dispositivo 202. La señal es dividida en tramas de tiempo con una duración de algunas decenas o centenares de milisegundos; por ejemplo, véase el número 502 y las líneas discontinuas. A continuación, la señal es analizada trama a trama utilizando, por ejemplo, un

análisis cepstral, durante el cual se calculan varios coeficientes cepstrales determinando una transformada de Fourier de la trama y rompiendo la correlación con una transformada de coseno para recoger los coeficientes dominantes, por ejemplo los 10 primeros coeficientes por trama. También pueden determinarse coeficientes derivados para estimar la dinámica 504 de la voz.

A continuación, el vector de características que comprende los coeficientes obtenidos y que representa la trama de voz es sometido a un clasificador acústico, por ejemplo un clasificador de red neural que asocie los vectores de características con diferentes fonemas 506; es decir, el vector de características es ligado a cada fonema con cierta probabilidad. El clasificador puede ser personalizado mediante configuraciones ajustables o procedimientos de entrenamiento expuestos en lo que antecede.

Acto seguido las secuencias de fonemas que pueden ser construidas concatenando los fonemas que posiblemente subyagan a los vectores de características son analizadas con un decodificador HMM (modelo oculto de Márkov) u otro decodificador adecuado que determine la trayectoria 508 del fonema más probable (y el correspondiente elemento de nivel superior, por ejemplo una palabra) (formando en la figura una frase “esto parece…”) a partir de las secuencias utilizando, por ejemplo, un léxico dependiente del contexto y/o un modelo gramatical de lenguaje y vocabulario relacionado. Tal trayectoria se denomina a menudo trayectoria de Viterbi, y maximiza la probabilidad a posteriori para la secuencia en relación con el modelo probabilístico dado.

Considerando el aspecto de compartición de tareas de la invención, la compartición 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 asignar, en base a parámetros/reglas predeterminados o a negociaciones dinámicas/de tiempo real, las tareas tras las etapas 502, 504, 506 y 508 de reconocimiento, de modo que el dispositivo 202 se ocupa de varias etapas (por ejemplo, 502), tras lo cual el servidor 208 ejecuta las restantes etapas (504, 506 y 508, respectivamente). De forma alternativa, el dispositivo 202 y el servidor 208 ejecutarán ambos todas las etapas, pero únicamente en relación con una porción de la señal de voz, en cuyo caso las porciones convertidas de voz a texto serán combinadas finalmente por el dispositivo 202, el servidor 208 o alguna otra entidad para establecer todo el texto. No obstante, en una alternativa, puede sacarse provecho de las dos opciones anteriores simultáneamente; por ejemplo, el dispositivo 202 se ocupa de al menos una tarea para toda la señal de voz (por ejemplo, la etapa 502) debido, por ejemplo, a un nivel de servicio actual que explícitamente así lo define, y también ejecuta las etapas restantes para una porción pequeña de la voz concurrente con la ejecución de las mismas etapas restantes para el resto de la voz por parte del servidor 208. Tal división flexible de tareas puede originarse en la optimización temporal del proceso global de conversión de voz a texto, es decir, se estima que, por la división aplicada, el dispositivo 202 y el servidor 208 terminarán sus tareas de forma sustancialmente simultánea y, así, el tiempo de respuesta percibido por el usuario del dispositivo 202 se minimiza desde el lado del servicio.

Los sistemas modernos de reconocimiento de voz pueden llegar a una tasa de reconocimiento de aproximadamente el 80% si la señal de voz de entrada es de buena calidad (libre de perturbaciones y ruido de fondo, etc.), pero la tasa puede descender hasta el 50% o así en condiciones más exigentes. Por lo tanto, tal como se ha presentado anteriormente, algún tipo de edición puede mejorar notablemente el rendimiento del motor de reconocimiento básico.

La Figura 6 da a conocer una opción de componentes básicos del dispositivo electrónico móvil 202 como un terminal móvil o una PDA dotada de capacidades de comunicaciones, ya sean internas o externas. La memoria 604, dividida entre uno o más chips físicos de memoria, comprende el código necesario, por ejemplo en forma de programa/aplicación informático 612 para permitir la edición de la voz y una conversión al menos parcial de voz a texto (~motor de reconocimiento de voz) y otros datos 610, por ejemplo las configuraciones actuales, voz de forma digital (opcionalmente codificada) y datos de reconocimiento de voz. La memoria 604 puede, además, referirse a una tarjeta de memoria preferentemente extraíble, un disquete, un CD-ROM o un medio fijo de almacenamiento, como un disco duro. La memoria 604 puede ser, por ejemplo, ROM o RAM por naturaleza. Para la ejecución real del código almacenado en la memoria 604 se requiere el medio 602 de procesamiento, por ejemplo una unidad de procesamiento/de control como un microprocesador, un DSP, un microcontrolador o un chip de lógica programable, que opcionalmente comprenda una pluralidad de (sub)unidades cooperantes o paralelas. La pantalla 606 y el teclado o la botonera 608 u otro medio aplicable de entrada de control (por ejemplo, una pantalla táctil o una entrada de control de voz) proporcionan al usuario del dispositivo 202 con un medio de control del dispositivo y de visualización de datos (~interfaz de usuario). El medio 616 de entrada de voz incluye un sensor/transductor, por ejemplo un micrófono y u convertidor A/D para recibir una señal de entrada acústica y para transformar la señal acústica recibida en una señal digital. Se requiere el medio inalámbrico de transferencia de datos 614, por ejemplo un transceptor de radio (GSM, UMTS, WLAN, Bluetooth, infrarrojo, etc.) para la comunicación con otros dispositivos.

La Figura 7 da a conocer un correspondiente diagrama de bloques del servidor 208. El servidor comprende una unidad 702 de control y una memoria 704. La unidad 702 de control para controlar el motor de reconocimiento de voz y otras funcionalidades del servidor 208, incluyendo el intercambio de la información de control, que puede, en la práctica tener lugar a través del medio 714/718 de entrada/salida de datos o de otro medio de comunicaciones, puede ser implementada como una unidad de procesamiento o una pluralidad de unidades cooperantes como el medio 602 de procesamiento del dispositivo electrónico móvil 202. La memoria 704 comprende la aplicación 712 del lado del servidor que debe ser ejecutada por la unidad 702 de control para llevar a cabo al menos algunas tareas del proceso global de conversión de voz a texto, por ejemplo un motor de reconocimiento de voz. Véase el párrafo anterior para ejemplos de

posibles implementaciones de la memoria. Pueden proporcionarse aplicaciones/procesos 716 opcionales para implementar servicios adicionales. Los datos 710 incluyen datos de voz, parámetros de reconocimiento de voz, configuraciones, etc. Es evidente que, al menos, parte de la información requerida puede ubicarse en unas instalaciones de almacenamiento remoto, por ejemplo una base de datos, a las que el servidor 808 tiene acceso, por ejemplo, a través del medio 714 de entrada de datos y del medio 718 de salida de datos. El medio 714 de entrada de datos comprende, por ejemplo, una interfaz o un adaptador de red (Ethernet, WLAN, Token Ring, ATM, etc.) para recibir datos de voz e información de control enviados por el dispositivo 202. Así mismo, se incluye el medio 718 de salida de datos para transmitir, por ejemplo, los resultados de la compartición de tareas a la siguiente etapa. En la práctica, el medio 714 de entrada de datos y el medio 718 de salida de datos pueden combinarse en una única interfaz multidireccional accesible por la unidad 702 de control.

El dispositivo 202 y el servidor 208 pueden ser realizados como una combinación de soporte lógico hecho a medida y un soporte físico más genérico o, de forma alternativa, por medio de un soporte físico más especializado, 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 proporcionado en un soporte como un disquete, un CD, un disco duro o una tarjeta de memoria.

El alcance de la presente invención puede encontrarse en las siguientes reivindicaciones. Sin embargo, los dispositivos utilizados, las etapas de los procedimientos, los detalles de la compartición de tareas, etc., pueden depender de un caso de uso particular que, pese a todo, converge en las ideas básicas presentadas en lo que antecede, tal como apreciará un lector experto en la técnica.

REIVINDICACIONES

1. Un dispositivo móvil operable en una red de comunicaciones inalámbricas que comprende:

- un medio de entrada de voz para recibir voz y convertir la voz en una señal digital (616) de voz representativa,
- un medio de entrada de control para comunicar una orden de edición relativa a la señal digital (608) de voz,
- un medio de procesamiento para llevar a cabo una tarea de edición de la señal digital de voz en respuesta a la orden (602) de edición recibida,
- al menos parte de un motor de reconocimiento de voz para llevar a cabo tareas de conversión (612) a texto de la señal digital de voz, y
- un transceptor para intercambiar información relativa a la señal digital de voz y la conversión de la misma de voz a texto con una entidad externa conectada funcionalmente a dicha red (614) de comunicaciones inalámbricas, en el que dicho dispositivo móvil está configurado opcionalmente para transmitir a otra entidad el texto resultante de la conversión de voz a texto para una tarea ulterior de procesamiento seleccionada del grupo constituido por: revisión ortográfica, traducción automática, traducción humana, verificación de la traducción y síntesis de texto a voz.

2. El dispositivo móvil según la reivindicación 1 que, además, comprende un medio de visualización para visualizar al menos parte de la señal digital de voz, tras lo cual dicho medio de entrada de control está configurado para comunicar una orden de edición relativa a dicha parte visualizada, en el que la visualización de la señal comprende opcionalmente al menos un elemento seleccionado del grupo constituido por: una representación del dominio temporal de la señal, una representación del dominio frecuencial de la señal, una parametrización de la señal, una operación de acercamiento y alejamiento dirigida a la señal visualizada, un valor numérico determinado a partir de una porción de la señal definida por el usuario, un puntero a una ubicación definida por el usuario en la señal visualizada, y el resalte de una subzona de la señal visualizada definida por el usuario, y en el que dicho dispositivo móvil está configurado opcionalmente, además, para visualizar al menos una porción del texto resultante de la conversión según alinea en relación con la porción correspondiente visualizada de la señal.

3. El dispositivo móvil según la reivindicación 1 en el que dicha tarea de edición está seleccionada del grupo constituido por: una supresión de una porción de la señal, una inserción de una porción de voz en la señal, la regrabación de una porción de la señal, la sustitución de una porción de la señal, el cambio en la amplitud de la señal, el cambio en el contenido espectral de la señal, el cambio de la dinámica de la señal y la ejecución de un algoritmo de reducción de ruido.

4. El dispositivo móvil según la reivindicación 1 en el que dicha al menos parte del motor de reconocimiento de voz comprende un elemento seleccionado del grupo constituido por: un preprocesador para dividir la señal digital de voz en tramas de una longitud predeterminada, un codificador de audio para comprimir la señal digital de voz, un analizador cepstral, un clasificador acústico, un clasificador de red neural, un decodificador de trayectoria óptima, un decodificador HMM (modelo oculto de Márkov), un modelo léxico de lenguaje, un modelo gramatical de lenguaje, un modelo léxico de lenguaje dependiente del contexto, un modelo gramatical de lenguaje dependiente del contexto, configuraciones específicas del usuario y vocabulario.

5. El dispositivo móvil según la reivindicación 1 en el que dicho información intercambiada incluye un elemento seleccionado del grupo constituido por: voz en forma digital, voz digital codificada, información de estado del dispositivo, reconocimiento de mensajes, información de control, orden de edición, datos de negociación de compartición de tareas, el valor de un parámetro relacionado con la compartición de tareas, estado de tareas, aviso de interrupción del servicio, cifra de carga, resultado intermedio de la conversión de voz a texto.

6. El dispositivo móvil según la reivindicación 1 en el que dicha información se intercambia utilizando al menos una práctica se comunicación seleccionada del grupo constituido por: un mensaje SMS (servicio de mensajes cortos), un mensaje MMS (servicio de mensajes multimedia), un correo electrónico, una llamada de datos, una conexión GPRS (servicio general de radiotransmisión por paquetes) y una llamada de voz.

7. El dispositivo móvil según la reivindicación 1 configurado para compartir la ejecución de tareas requeridas para llevar a cabo la conversión de voz a texto con la entidad externa, estando además configurado dicho dispositivo móvil, opcionalmente, para compartir la ejecución de tareas para optimizar un factor según criterios predeterminados, seleccionándose dicho factor del grupo constituido por: tiempo de ejecución de la conversión de voz a texto, costo de la conversión, cantidad de la transferencia de datos requerida, carga de procesamiento y carga de memoria.

8. El dispositivo móvil según la reivindicación 7 en el que la información intercambiada incluye al menos un elemento del grupo constituido por: data para asignar o llevar a cabo las tareas de la conversión de voz a texto, carga de procesamiento, carga de memoria, un estado de batería, una capacidad de batería, información sobre las tareas que se ejecutan con prioridad más elevada, ancho de banda disponible para la transmisión, tasa de transmisión de

los datos, costo del uso de la entidad externa por tamaño o duración de los datos de voz, tamaño o duración de la señal digital de voz, procedimiento disponible de codificación/decodificación, estado de la conversión, estado de la tarea, aviso de falta de disponibilidad del dispositivo, resultado intermedio de la conversión de voz a texto, voz digital, voz digital codificada, parámetro del reconocimiento de voz, y texto.

9. El dispositivo móvil según la reivindicación 7 configurado para utilizar resultados intermedios de la conversión de voz a texto, proporcionados tanto por el dispositivo como por la entidad externa para producir el texto.

10. El dispositivo móvil según la reivindicación 7 configurado para transmitir resultados intermedios de la conversión de voz a texto a otra entidad, por ejemplo a dicha entidad externa, para habilitar a otra entidad para que lleve a cabo al menos una de las opciones siguientes: combinar los resultados intermedios adquiridos del dispositivo móvil con los resultados obtenidos localmente para producir el texto, someter los resultados intermedios a un procesamiento adicional para producir el texto.

11. Un servidor operable en una red de comunicaciones que comprende:

- un medio de entrada de datos para recibir una señal digital de datos enviada por un dispositivo móvil, representando dicha señal digital de datos voz o, al menos, parte de la misma (714), y para recibir una orden de edición de la voz a través del dispositivo móvil,
- al menos parte de un motor de reconocimiento de voz para llevar a cabo tareas de conversión (712) a texto de una señal digital de datos,
- una unidad de control para intercambiar información de control con el dispositivo móvil, llevando a cabo una tarea de edición de la señal digital de voz en respuesta a la orden de edición recibida, y para determinar, en base a la información de control, las tareas que deben llevarse a cabo en la señal digital de datos recibida por dicha al menos parte del motor (702) de reconocimiento de voz, y
- un medio de salida de datos para comunicar al menos parte de la salida de las tareas llevadas a cabo a una entidad externa (718).

12. Un sistema para convertir voz a texto que comprende un dispositivo móvil (202) operable en una red de comunicaciones inalámbricas y un servidor (208) conectado funcionalmente a dicha red de comunicaciones inalámbricas, en el que:

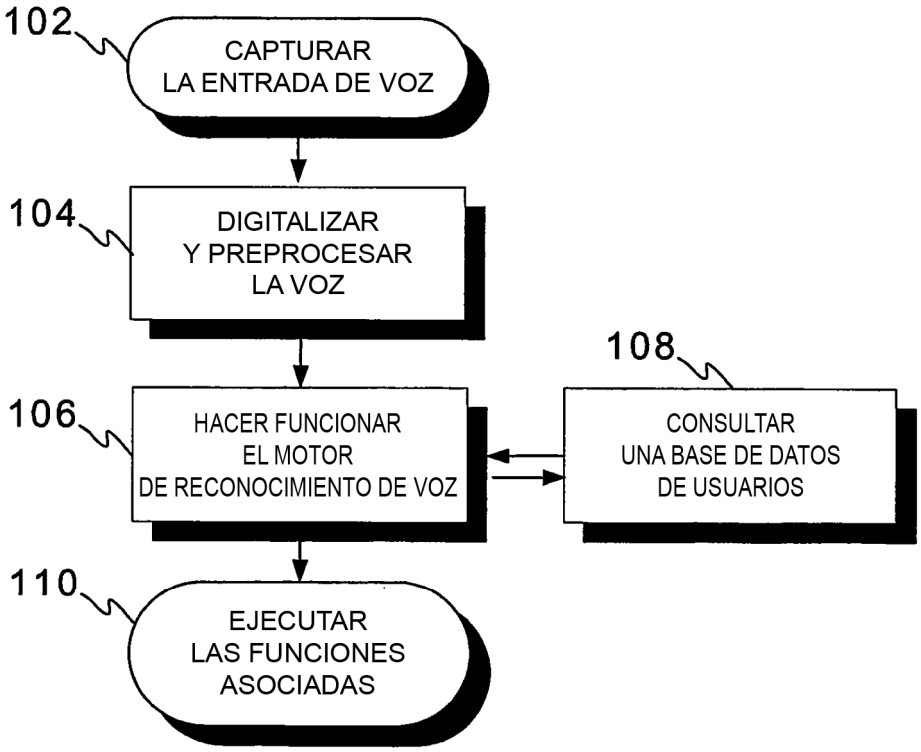
- dicho dispositivo móvil (202) está configurado para recibir voz y convertir la voz en una señal digital de voz representativa, para recibir una orden de edición relativa a la señal digital de voz, para procesar la señal digital de voz según la orden de edición, para intercambiar información relativa a la señal digital de voz y la conversión de voz a texto de la misma con el servidor (208), y para ejecutar parte de las tareas requeridas para llevar a cabo una conversión a texto de la señal digital de voz, y
- dicho servidor (208) está configurado para recibir información relativa a la señal digital de voz y la conversión de voz a texto de la misma, y para ejecutar, en base a la información intercambiada, la parte restante de las tareas requeridas para llevar a cabo una conversión a texto de la señal digital de voz.

13. Un procedimiento para convertir voz en texto que tiene las etapas de:

- recibir, en un dispositivo móvil operable en una red inalámbrica, una fuente de voz, y convertir la fuente de voz en una señal digital (304) de voz representativa,
- recibir una orden de edición relativa a la señal digital (308) de voz por parte del dispositivo móvil,
- procesar la señal digital de voz según la orden (310) de edición,
- intercambiar información relativa a la señal digital de voz y la conversión de voz a texto de la misma (312), y
- ejecutar, en base a la información intercambiada, al menos parte de las tareas requeridas para llevar a cabo una conversión de voz a texto de la señal digital (314) de voz, comprendiendo el procedimiento además, opcionalmente, visualizar al menos parte de la señal digital de voz en una pantalla del dispositivo móvil, tras lo cual la orden de edición recibida se relaciona adicionalmente, de forma opcional, con dicha parte visualizada.

14. Un programa ejecutable en ordenador que comprende medios de código adaptados, cuando se ejecutan en un ordenador, para llevar a cabo las etapas del procedimiento tal como se definen en la reivindicación 13.

15. Un medio de soporte que comprende el programa ejecutable en ordenador de la reivindicación 14, en el que dicho medio de soporte incluye opcionalmente al menos un elemento seleccionado del grupo constituido por: una tarjeta de memoria, un disquete flexible, un CD-ROM y un disco duro.



TÉCNICA ANTERIOR

Figura 1

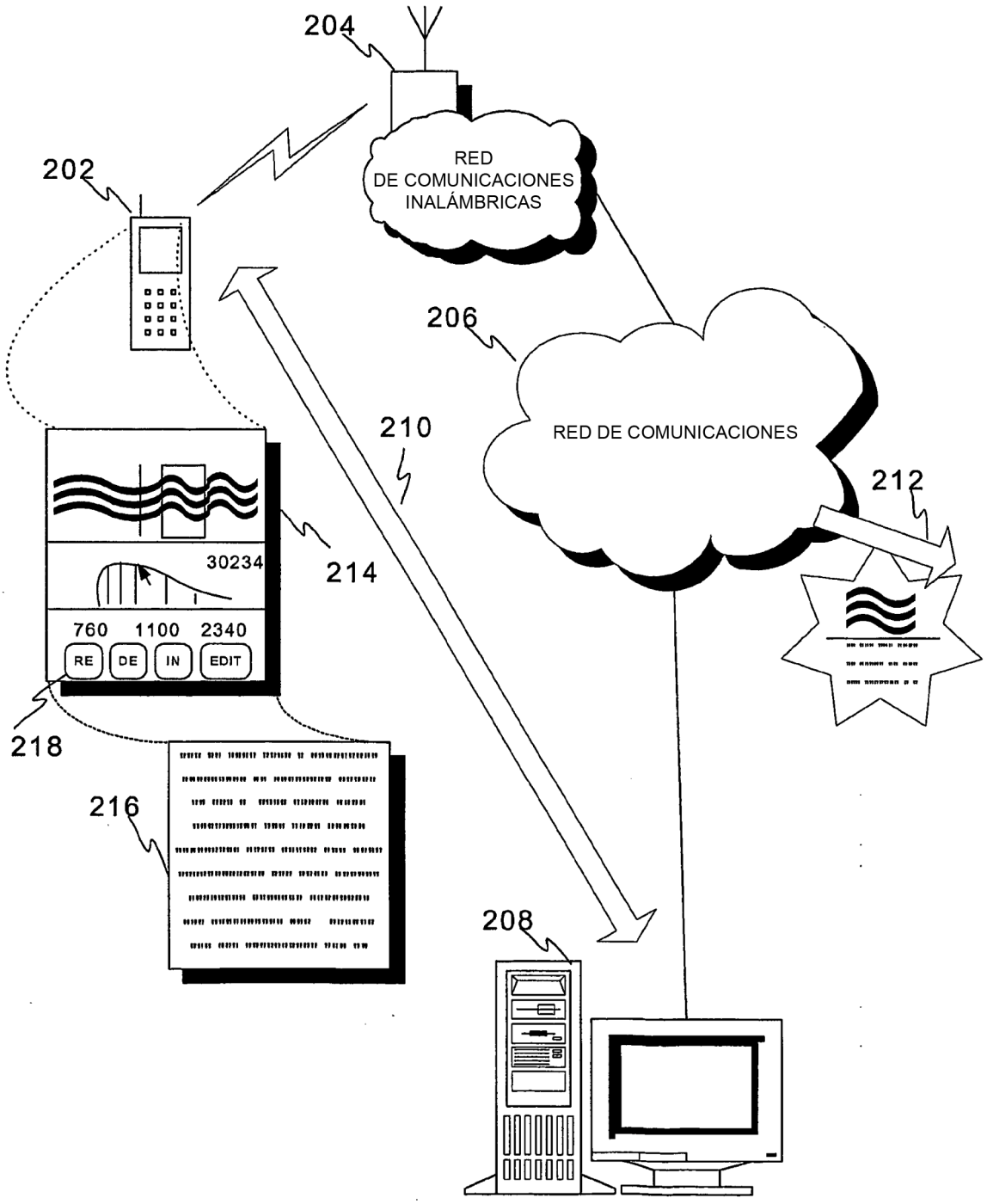

204
202
RED
DE COMUNICACIONES
INALÁMBRICAS
206
RED DE COMUNICACIONES
210
212
214
30234
760 1100 2340
RE DE IN EDIT
218
216
208

Figura 2

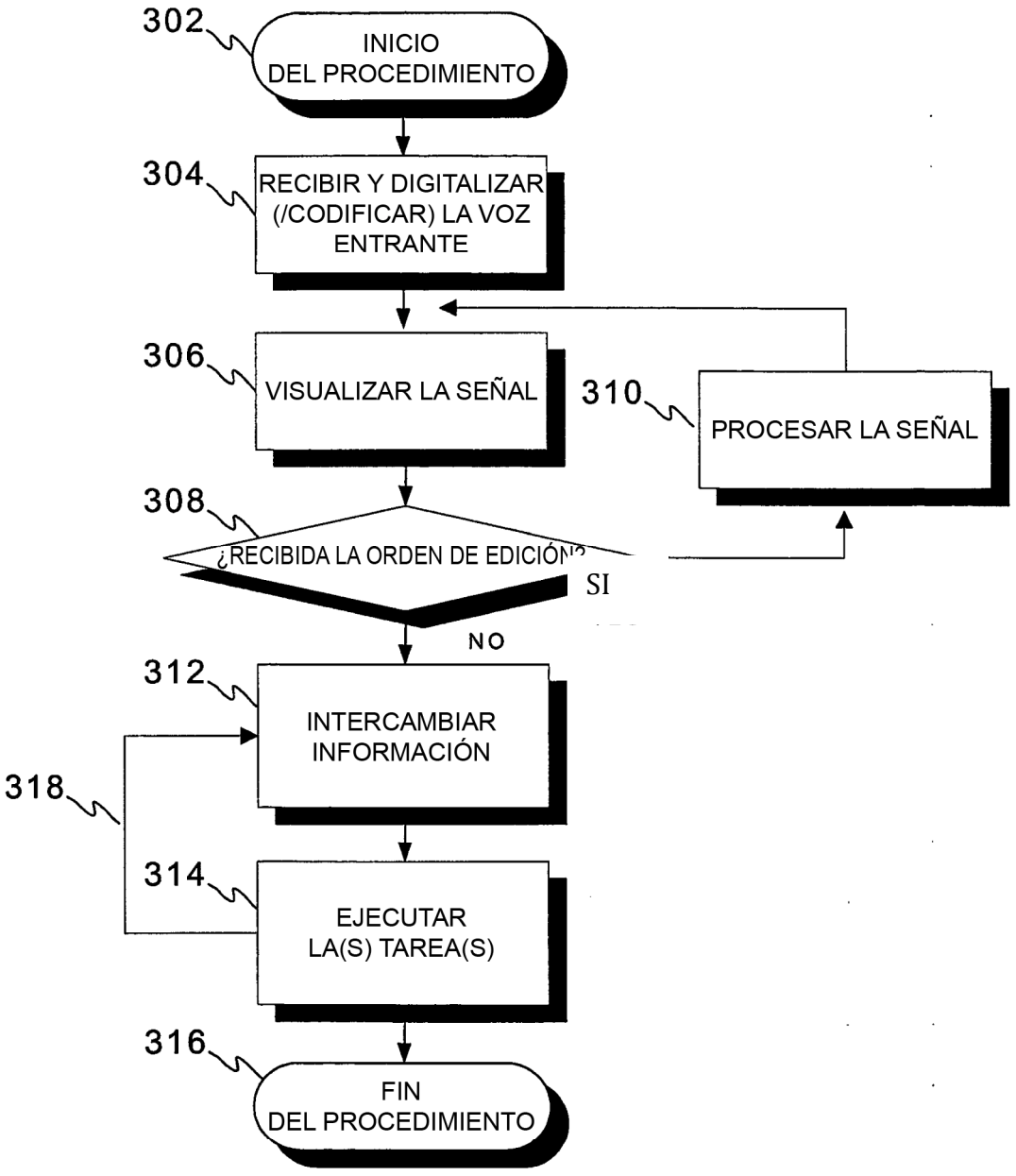

302
INICIO
DEL PROCEDIMIENTO
304
RECIBIR Y DIGITALIZAR
(/CODIFICAR) LA VOZ
ENTRANTE
306
VISUALIZAR LA SEÑAL
310
PROCESAR LA SEÑAL
308
¿RECIBIDA LA ORDEN DE EDICIÓN?
SI
NO
312
INTERCAMBIAR
INFORMACIÓN
318
314
EJECUTAR
LA(S) TAREA(S)
316
FIN
DEL PROCEDIMIENTO

Figura 3

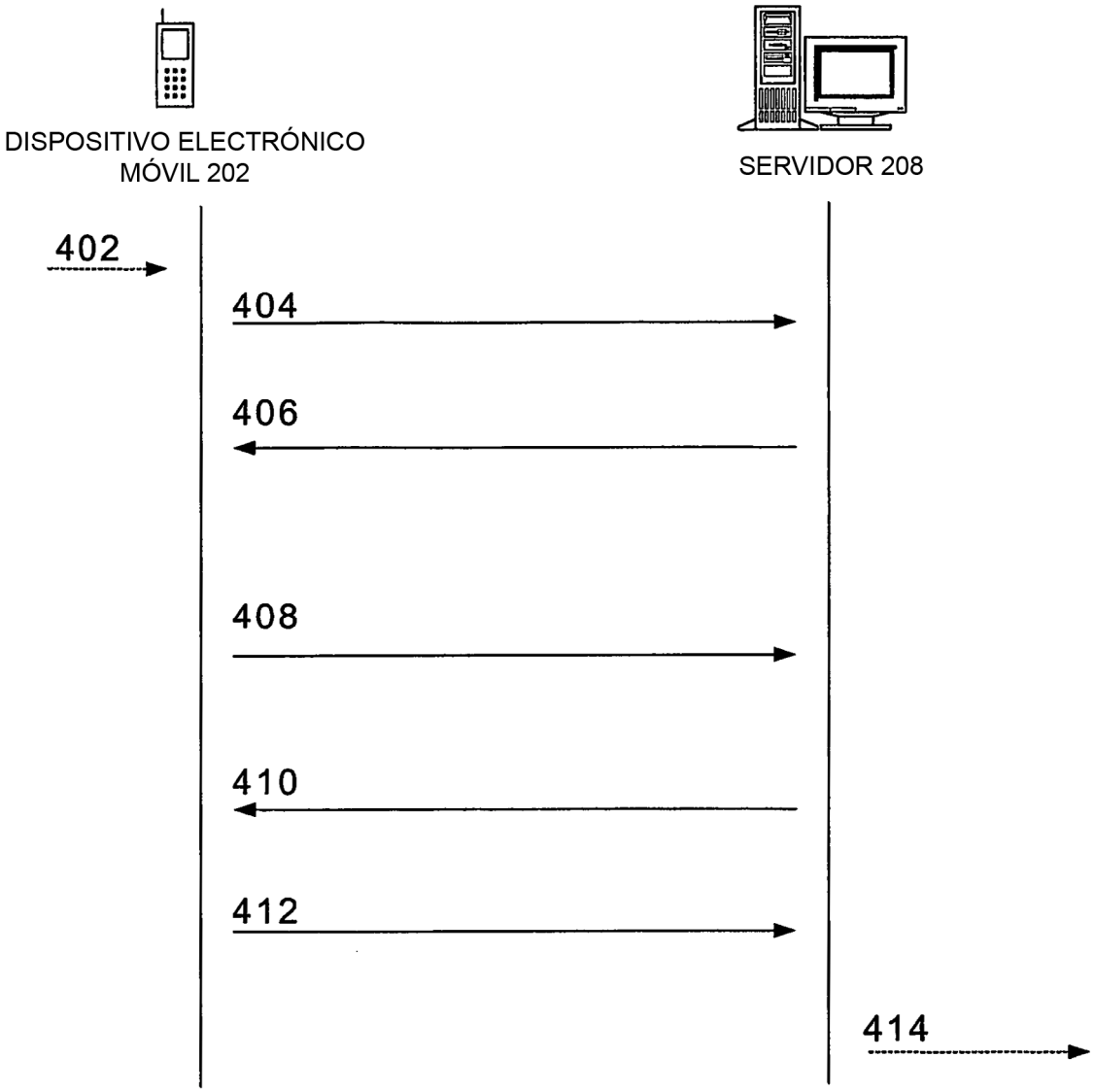

DISPOSITIVO ELECTRÓNICO MÓVIL 202
SERVIDOR 208
402
404
406
408
410
412
414

Figura 4

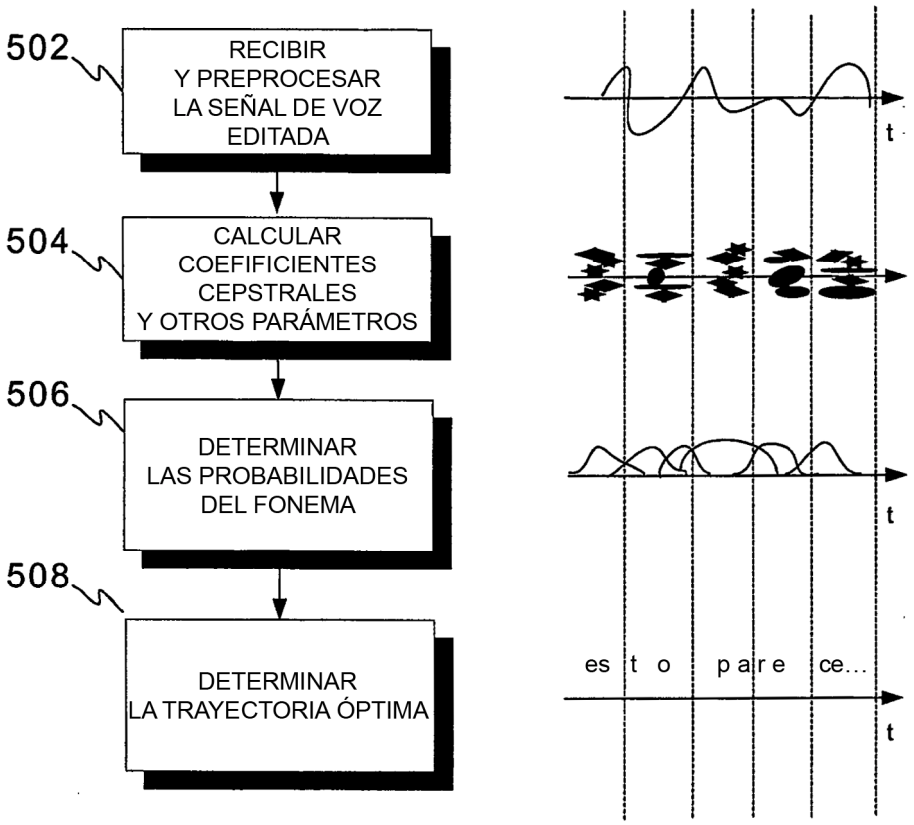

502
RECIBIR
Y PREPROCESAR
LA SEÑAL DE VOZ
EDITADA
504
CALCULAR
COEFIFICIENTES
CEPSTRALES
Y OTROS PARÁMETROS
506
DETERMINAR
LAS PROBABILIDADES
DEL FONEMA
508
DETERMINAR
LA TRAYECTORIA ÓPTIMA
t
t
es t o p a i r e ce...
t

Figura 5

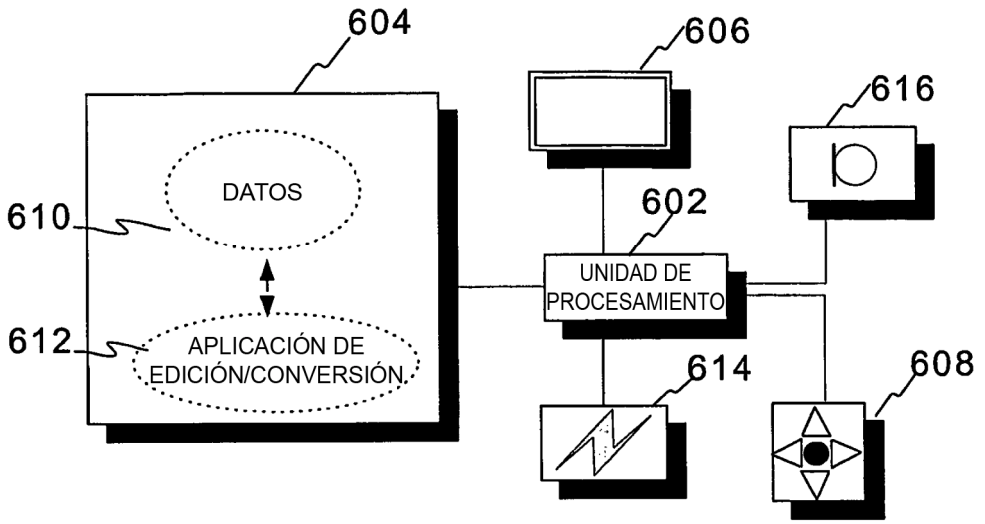


Figura 6

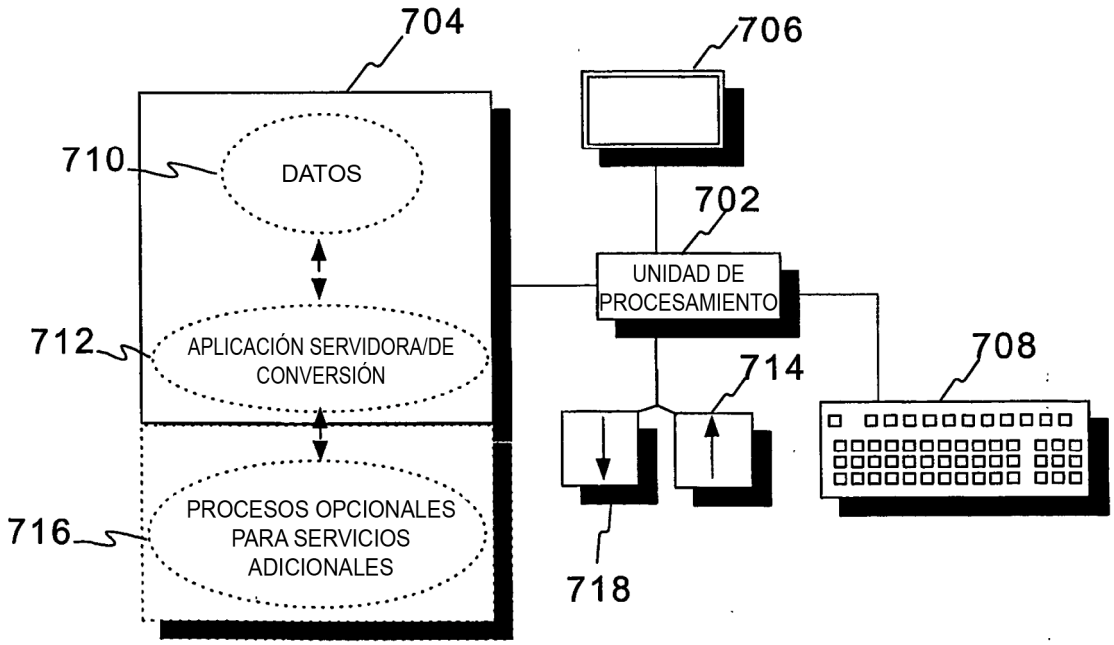


Figura 7