(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 375 537**

(51) Int. CI.:

H04L 29/08	(2006.01)
G09B 21/00	(2006.01)
H04M 3/42	(2006.01)
G10L 15/26	(2006.01)
G09B 21/04	(2006.01)
H04N 7/14	(2006.01)
H04N 7/15	(2006.01)
H04L 29/06	(2006.01)
G06F 17/30	(2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **08706572 .8**

(96) Fecha de presentación: **28.01.2008**

(97) Número de publicación de la solicitud: **2154885**

(97) Fecha de publicación de la solicitud: **17.02.2010**

(54) Título: **UN MÉTODO DE VISUALIZACIÓN DE SUBTÍTULOS Y UN DISPOSITIVO DE CONTROL DE VIDEOCOMUNICACIÓN.**

(30) Prioridad:
17.05.2007 CN 200710074542

(45) Fecha de publicación de la mención BOPI:
01.03.2012

(45) Fecha de la publicación del folleto de la patente:
01.03.2012

(73) Titular/es:
Huawei Technologies Co., Ltd.
Huawei Administration Building Bantian
Longgang District, Shenzhen
Guangdong 518129 , CN

(72) Inventor/es:
LIU, Zhihui y
YUE, Zhonghui

(74) Agente: **Lehmann Novo, Isabel**

DESCRIPCIÓN

Un método de visualización de subtítulos y un dispositivo de control de videocomunicación

ANTECEDENTES DE LA INVENCIÓN

CAMPO DE LA TECNOLOGÍA

La presente invención se refiere a un campo de la comunicación y más en particular, a un método de visualización de subtítulos y un sistema y dispositivo de videocomunicación.

ANTECEDENTES DE LA INVENCIÓN

Con el desarrollo de tecnologías tales como Protocolo de Voz sobre Internet (IP) (VoIP), Procesamiento Digital de Señales (DSP) y ancho de banda de red, los usuarios pueden ahora efectuar, de forma cómoda, llamadas a larga distancia a través de un sistema de videoconferencia y ver las expresiones y acciones de la parte opuesta a través de imágenes. Un sistema de videoconferencia convencional suele incluir terminales de vídeo, una red de transmisión y una unidad de control multipunto (MCU). El terminal de vídeo está adaptado para recibir una entrada de audio y vídeo local, codificar un flujo de códigos de audio y de vídeo y enviar el flujo de códigos de audio y de vídeo codificado a un extremo lejano a través de la red de transmisión. El terminal de vídeo está también adaptado para recibir y decodificar el flujo de códigos de audio y de vídeo desde el extremo lejano y para recuperar el sonido y las imágenes del extremo lejano a través de altavoces y monitores locales. La unidad MCU está principalmente adaptada para realizar un control de conferencias y un intercambio de medios, por ejemplo, una o más conferencias multipunto se pueden celebrar a través de la MCU. La MCU recibe, además, los flujos de códigos de audio y de vídeo desde los terminales, realiza una mezcla de audio, combina múltiples tramas y envía los flujos de códigos de audio y de vídeo procesados a los terminales correspondientes. Sin embargo, el sistema de videocomunicación convencional, en su mayor parte, no soporta una función de visualización de subtítulos en tiempo real. Si el subtítulo se visualiza en tiempo real durante la conferencia, el sonido de un altavoz se convierte en caracteres y los caracteres se visualizan, de forma síncrona, en una pantalla, se mejora la experiencia visual de los usuarios y las personas con dificultades auditivas o barrera de lenguaje pueden realizar la comunicación utilizando el sistema de vídeo con el fin de superar la dificultad de comunicación.

En una solución existente de la videoconferencia con la presentación visual de subtítulos, durante una conferencia, el subtítulo que necesita enviarse se aplica a la entrada del terminal a través de una interfaz, siendo el texto de entrada superpuesto sobre la imagen y dicha imagen se codifica y envía a un extremo receptor. Después de recibir la imagen, el extremo receptor decodifica y visualiza la imagen. Sin embargo, el defecto de la técnica anterior es que se requiere una entrada manual excesiva, por lo que el contenido de los subtítulos, que se van a visualizar, debe editarse por anticipado y el contenido de los subtítulos no se puede transmitir en tiempo real, por lo que el método de visualización de subtítulos sólo suele ser aplicable a la notificación de información de la conferencia.

La patente de Estados Unidos nº 5.774.857 da a conocer un dispositivo de comunicación con visualización de subtítulos. Un teléfono y algunas interfaces de usuarios, por ejemplo, un teclado, están exteriormente conectados a un extremo receptor y un módulo de reconocimiento de voz está integrado. Cuando un usuario establece una llamada a través del teléfono, el dispositivo recibe señales de voz desde un extremo lejano y a continuación, un módulo de reconocimiento de voz convierte las señales de voz recibidas en señales de texto, un modulador de radiofrecuencia (RF) modula las señales de texto a señales de banda base de vídeo y envía las señales a un aparato de TV para su presentación visual.

El documento US 7.039.675 B1 está relacionado con el reconocimiento de voz simple en una videoconferencia. Da a conocer un método y un aparato para un terminal de uso general que se conecta a un sistema de videoconferencia multipunto, a través de Internet, para participar en la sesión de videoconferencia. Se centra en el reconocimiento y conversión de las señales de audio mezcladas en señales de texto y para enviar las señales de texto al terminal de uso general.

El documento US 2004/119814 A1 da a conocer un terminal de videoconferencia lejano para participar en una conferencia a través de enlaces de comunicaciones inalámbricos. Además, da a conocer que el dispositivo del cliente puede recibir texto que acompaña al vídeo y multiplexar el texto con el vídeo para proporcionar una señal de vídeo cerrada con subtítulos.

El documento CN 1 283 931 A da a conocer un método de superposición de subtítulos.

Willebeek-Lemair M H et al: “Unidades de control multipunto para sesiones de videoconferencias”, Processing of the Conference on Local Computer Networks, 2 de octubre de 1994, páginas 356–364, da a conocer una infraestructura de unidad de control multipunto para sesiones de videoconferencias.

Clark W J: “Conferencias multimedia multipunto”, IEEE Communications Magazine, IEEE Service Center, Piscataway, Estados Unidos, tomo 30, nº 5, 1 de mayo de 1992, páginas 44–50, da a conocer un método para sesiones de conferencia multimedia multipunto y la función de mezcla de audio se puede poner en práctica en la MCU.

Sin embargo, durante la puesta en práctica de la presente invención, los inventores encuentran al menos los problemas técnicos siguientes.

En el dispositivo existente, el modulador de RF modula las señales de texto en señales de RF y modula las señales de RF para la presentación visual de señales de banda base de vídeo, por lo que se aumenta, en gran medida, la complejidad de la visualización de los subtítulos y se tiene un rendimiento en tiempo real no favorable. A continuación, el módulo de reconocimiento de voz del dispositivo está dispuesto en el extremo receptor, lo que no es conveniente para la formación en reconocimiento de voz de los usuarios. Además, en la conferencia multipunto, si las señales de voz recibidas por el dispositivo existente son señales sintetizadas de voz de una pluralidad de personas, el módulo de reconocimiento de voz único no puede reconocer las diferentes señales de voz al mismo tiempo, por lo que las señales de reconocimiento están desordenadas y el subtítulo no se puede visualizar de forma correcta.

SUMARIO DE LA INVENCIÓN

Con el fin de resolver los problemas de un sistema de videocomunicación existente en el sentido de que una tecnología de visualización de subtítulos es complicada y su rendimiento en tiempo real es desfavorable, la presente invención da a conocer un método de visualización de subtítulos y un dispositivo de control de videocomunicación según las reivindicaciones independientes 1 y 2, respectivamente. El método de visualización de subtítulos es simple y el rendimiento en tiempo real es alto.

En comparación con la técnica anterior, las soluciones técnicas en las formas de realización de la presente invención presentan al menos las ventajas siguientes. En las formas de realización de la presente invención, las señales de voz se reconocen para las señales de texto y las señales de texto son directamente superpuestas sobre las señales de vídeo para su codificación y transmisión, de modo que los usuarios puedan decodificar directamente y visualizar imágenes e información de caracteres correspondientes a una voz y el método es simple y el rendimiento en tiempo real es alto.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Para ilustrar las soluciones técnicas, según las formas de realización de la presente invención, o en la técnica anterior, con mayor claridad, se hace referencia, a continuación, a los dibujos adjuntos para describir las formas de realización o la técnica anterior, de una forma concisa.

La Figura 1 es una vista esquemática simple de principios de una comunicación punto a punto de un sistema y dispositivo de videocomunicación según una primera forma de realización de la presente invención;

La Figura 2 es una vista esquemática simple de principios de una comunicación punto a punto de un sistema y dispositivo de videocomunicación según una segunda forma de realización de la presente invención;

La Figura 3 es una vista esquemática simple de principios de una comunicación multipunto de un sistema y dispositivo de videocomunicación, según una forma de realización de la presente invención y

La Figura 4 es un diagrama de flujo esquemático simple de un método de visualización de subtítulos en una videocomunicación, según una forma de realización de la presente invención.

DESCRIPCIÓN DETALLADA DE LAS FORMAS DE REALIZACIÓN

Las soluciones técnicas según las formas de realización de la presente invención se describen, de forma clara y completa, haciendo referencia a los dibujos adjuntos de las formas de realización. Evidentemente, las formas de realización aquí ilustradas son solamente una parte, y no la totalidad, de las formas de realización de la presente invención.

En la presente invención, un sistema de videocomunicación comprende terminales de vídeo, módulos de reconocimiento de voz, módulos de codificación de vídeo, una unidad MCU y una red de transmisión, en donde las señales de voz introducidas por los usuarios y las señales de vídeo a través de los terminales de vídeo, son convertidas por los módulos de reconocimiento de voz desde las señales de voz de entrada a señales de texto, mientras que los módulos de codificación de vídeo realizan una superposición de los subtítulos sobre las señales de texto y las señales de vídeo y codifican las señales de texto y las señales de vídeo y asimismo, envían las señales de texto y las señales de vídeo a un extremo lejano a través de la red de transmisión.

Haciendo referencia a la Figura 1, un módulo de reconocimiento de voz 10 y un módulo de codificación de vídeo 20, según una primera forma de realización de la presente invención, están integrados en un terminal de vídeo. El módulo de reconocimiento de voz 10 está conectado a un módulo de captura de voz (micrófonos) y está adaptado para reconocer las señales de voz recogidas por el micrófono en señales de texto y para transmitir las señales de texto al módulo de codificación de vídeo 20. El módulo de codificación de vídeo 20 está conectado a una cámara de vídeo y está adaptado para superponer las señales de texto sobre señales de vídeo de imágenes recogidas por un módulo de captura de imagen (cámara de vídeo), para codificar las señales de texto y las señales de vídeo de imágenes y para enviar las

señales de texto y las señales de vídeo de imagen a un extremo lejano, de modo que los usuarios lejanos puedan ver la información de subtítulos reconocida en una presentación visual síncrona con las señales de voz en tiempo real, con lo que se mejora la experiencia en sesiones de los usuarios y en particular, algunas personas con dificultades auditivas pueden establecer una comunicación normal.

Conviene señalar que el módulo de reconocimiento de voz 10 de esta forma de realización está integrado en el terminal de vídeo, de modo que resulta cómodo para el usuario realizar una formación en reconocimiento de voz para mejorar una tasa de reconocimiento. A continuación, el módulo de reconocimiento de voz 10 se puede configurar de forma que sea activado en función de las demandas del usuario. Cuando múltiples usuarios participan en una conferencia multipunto, se soporta una función de distribución multipunto. Además, un módulo sintetizador de voz se puede integrar, además, en el terminal de vídeo de esta forma de realización. Después de que el usuario introduzca información de caracteres a través de un módulo de entrada de caracteres (un teclado o en otros modos), el módulo sintetizador de voz convierte la información de caracteres en señales de voz y envía la información de caracteres al módulo de codificación de vídeo 20. El módulo de codificación de vídeo 20 superpone la entrada de información por el usuario sobre las señales de vídeo de imágenes, codifica la información y las señales de vídeo de imágenes y envía la información y las señales de vídeo de imágenes al extremo lejano. Al mismo tiempo, las señales de voz convertidas se envían a un módulo de codificación de voz para su codificación y se envían al extremo lejano. De este modo, una persona sorda o una persona con barrera de lenguaje pueden comunicarse con una parte opuesta a través del dispositivo de videocomunicación según la presente invención.

Haciendo referencia a la Figura 2, los módulos de reconocimiento de voz y los módulos de codificación de vídeo de una segunda forma de realización de la presente invención están integrados en una MCU. Una pluralidad de módulos de reconocimiento de voz y de módulos de codificación de vídeo están integrados en la MCU. En este caso, los terminales de comunicación realizan el control de la conferencia y el intercambio de medios a través de la unidad MCU. En correspondencia, la MCU configura e inicia operativamente la pluralidad de módulos de reconocimiento de voz y de módulos de codificación de vídeo en función del número de usuarios que participan en la videocomunicación. Por ejemplo, en una conferencia punto a punto, cuando se reciben mensajes de voz de un terminal 1 y de un terminal 2, la MCU realiza un proceso de decodificación y luego, envía una señal de voz decodificada del terminal 1 a un primer módulo de reconocimiento de voz 11. El primer módulo de reconocimiento de voz 11 reconoce y convierte un sonido del terminal 1 en una señal de texto y transmite la señal de texto a un primer módulo de codificación de vídeo 21 correspondiente al terminal 2. El primer módulo de codificación de vídeo 21 superpone la señal de texto del terminal 1 sobre una imagen de vídeo, codifica la señal de texto y la imagen de vídeo y envía la señal de texto y la imagen de vídeo al terminal 2. Una señal de voz decodificada del terminal 2 se envía a un segundo módulo de reconocimiento de voz 12. El segundo módulo de reconocimiento de voz 12 reconoce y convierte un sonido del terminal 2 en una señal de texto y transmite la señal de texto a un segundo módulo de codificación de vídeo 22 correspondiente al terminal 1. El segundo módulo de codificación de vídeo 22 superpone la señal de texto del terminal 2 sobre una imagen de vídeo, codifica la señal de texto y la imagen de vídeo y envía la señal de texto y la imagen de vídeo al terminal 1. De este modo, después de decodificar respectivamente los flujos de códigos de vídeo recibidos, el terminal 1 y el terminal 2 pueden ver el subtítulo correspondiente.

Haciendo referencia a la Figura 3 en conjunto, cuando se celebra una conferencia multipunto a través de la segunda forma de realización de la presente invención, la unidad MCU configura e inicia, en forma correspondiente, la pluralidad de módulos de reconocimiento de voz y de módulos de codificación de vídeo, en función del número de usuarios que toman parte en la videocomunicación o del número de los módulos de reconocimiento de voz que se han de iniciar según se establece en el sistema. Por ejemplo, se inician operativamente tres módulos de reconocimiento de voz. En primer lugar, la MCU recibe datos de audio y de vídeo de cada terminal y decodifica los datos de audio. A continuación, un módulo de mezcla de audio realiza la mezcla de audio y, durante el proceso de mezcla de audio, se obtienen tres sitios operativos con el volumen máximo, por ejemplo, los sitios 1, 2 y 3. Los datos de voz de los tres sitios operativos, con el volumen máximo, se envían, respectivamente, a tres módulos de reconocimiento de voz. Los módulos de reconocimiento de voz reconocen las voces procedentes de los tres sitios operativos con el volumen máximo con el fin de obtener señales de texto T1, T2 y T3 en correspondencia con las voces y luego, envía las señales de texto a los módulos de codificación de vídeo correspondientes a los sitios operativos. Los módulos de codificación de vídeo superponen las señales de texto y las señales de vídeo y a continuación, envían las señales de texto y las señales de vídeo a los terminales. Por ejemplo, el subtítulo enviado al sitio operativo 1 es las señales de texto de voz después de que se hayan reconocido las voces del sitio 2 y del sitio 3, estando el subtítulo enviado al sitio 2 constituido por las señales de texto después de que se reconozcan las voces del sitio 1 y del sitio 3 y el subtítulo enviado al sitio 3 es las señales de texto después de que se reconozcan las voces del sitio 1 y del sitio 2. Los subtítulos enviados a otros sitios operativos son las señales de texto después de que se reconozcan las voces del sitio 1, del sitio 2 y del sitio 3.

Puede entenderse que, durante el proceso de mezcla de audio, la unidad MCU obtiene, en primer lugar, los tres sitios operativos 1, 2 y 3 con el sonido máximo y luego, superpone y codifica los sonidos de los sitios 2 y 3 y envía los sonidos al sitio 1. De este modo, las personas ubicadas en el sitio 1 oyen los sonidos del sitio 2 y del sitio 3. Los sonidos del sitio 1 y del sitio 3 son superpuestos y codificados y se envían al sitio 2. De este modo, las personas ubicadas en el sitio 2 pueden oír los sonidos del sitio 1 y del sitio 3. En correspondencia, las personas ubicadas en el sitio 3 pueden oír los sonidos de los sitios 1 y 2. Como tal, el subtítulo visualizado en cada sitio operativo está en correspondencia con el sonido oído por las personas en ese sitio. Después de recibir los flujos de códigos de audio y de vídeo, los terminales de

cada sitio operativo decodifican los flujos de códigos de audio y de vídeo, se puede recuperar el sonido y el subtítulo correspondiente al sonido se puede ver a través de los monitores y oírse por los altavoces. Además, la visualización de los subtítulos de la mezcla de audio de tres partes se toma como ejemplo en la descripción anterior; sin embargo, el dispositivo de videocomunicación de la presente invención se puede configurar para visualizar solamente el subtítulo de una parte con el volumen máximo o los subtítulos de dos partes con el volumen máximo o se pueden configurar para realizar la visualización de los subtítulos en otros modos de mezcla de audio multiparte, en función de la demanda de los usuarios.

Haciendo referencia a la Figura 4, según se describió anteriormente, un método de visualización de subtítulos, en una videocomunicación, según la presente invención, comprende las etapas siguientes.

En la etapa 1, se establece una videocomunicación.

En la etapa 2, se determina y establece una cantidad de señales de voz reconocidas.

En la etapa 3, se determina un volumen de cada sitio operativo y se seleccionan las señales de voz de altavoces con volumen máximo de un número correspondiente a dicha magnitud.

En la etapa 4, las señales de voz del altavoz se reconocen y convierten en señales de texto.

En la etapa 5, las señales de texto y las señales de vídeo de imagen, que necesitan recibirse por, y visualizarse para, otros participantes en la conferencia (señales de vídeo de imágenes de sitios operativos de los altavoces correspondientes a las señales de voz en esta forma de realización) se superponen y codifican en tiempo real y se envían, respectivamente, a otros participantes de la conferencia.

En la etapa 6, otros participantes en la conferencia reciben y decodifican las señales de vídeo con las señales de texto superpuestas y ven imágenes y subtítulos.

Se entiende que en el método, la cantidad de las señales de voz reconocidas se puede determinar y seleccionar en función de un valor establecido por el sistema o por medios manuales o la cantidad de las señales de voz reconocidas no se puede establecer y se reconoce la voz del sitio operativo de cada participante. A continuación, otros participantes en la conferencia pueden controlar y seleccionar, respectivamente, las imágenes del sitio necesarias a recibirse y visualizarse, pudiendo los participantes seleccionar la visualización de la imagen del sitio del hablante o las imágenes del sitio de otros no hablantes. Sin importar qué sitio se seleccione, sólo es necesario superponer y codificar la señal de texto del sitio hablante y la señal de vídeo de la imagen del sitio operativo a visualizarse, seleccionada por cada participante en la conferencia.

Se entiende que los módulos de reconocimiento de voz y los módulos de codificación de vídeo del sistema de videocomunicación de la presente invención se pueden disponer, además, en otros dispositivos o dispositivos dedicados en el sistema y la red de transmisión en conjunto o se pueden disponer, de forma separada, en dispositivos diferentes en el sistema y la red de transmisión. Los módulos de reconocimiento de voz y los módulos de codificación de vídeo cooperan para reconocer las señales de voz para las señales de texto y superponen directamente las señales de texto sobre las señales de vídeo para ser codificadas y transmitidas. El usuario puede decodificar directamente las imágenes de visualización y la información de caracteres correspondientes a la voz. Por lo tanto, el método es simple y el rendimiento en tiempo real es alto.

REIVINDICACIONES

1. Un método de visualización de subtítulos procesado por una Unidad de Control Multipunto, MCU, en una conferencia multipunto, comprendiendo la unidad MCU una pluralidad de módulos de reconocimiento de voz y una pluralidad de módulos de codificación de vídeo, configurando e iniciando la unidad MCU la pluralidad de módulos de reconocimiento de voz y de módulos de codificación de vídeo, en donde el número de módulos de reconocimiento de voz a iniciarse se establece en el sistema y comprendiendo el método las etapas que consisten en:

establecer una videocomunicación entre sitios operativos;

recibir, por medio de la MCU, señales de voz y señales de vídeo de cada sitio operativo y decodificar las señales de voz;

realizar, por la MCU, una mezcla de audio y obtener un número de los sitios operativos con volumen máximo durante el proceso de mezcla de audio, en donde el número de los sitios operativos con máximo volumen corresponde al número de los módulos de reconocimiento de voz iniciados por la MCU;

reconocer y convertir, respectivamente, por el número de los módulos de reconocimiento de voz, las señales de voz de los sitios operativos con el máximo volumen para las señales de texto correspondientes y enviar las señales de texto a los módulos de codificación de vídeo correspondientes a los sitios operativos;

superponer y codificar, por los módulos de codificación de vídeo, las señales de texto y las correspondientes señales de vídeo, que necesitan recibirse por, y visualizarse para, otros sitios operativos de conferencia y enviar las señales de texto, superpuestas y codificadas, y las señales de vídeo a los sitios operativos a través de la videocomunicación.

2. Un dispositivo de control de videocomunicación, que comprende una unidad de control multipunto, MCU, y una pluralidad de dispositivos de terminales de vídeo, estando la pluralidad de dispositivos terminales de vídeo conectados a la MCU, comprendiendo cada uno de los dispositivos terminales de vídeo un módulo de captura de voz y un módulo de captura de imagen, integrando la MCU una pluralidad de módulos de reconocimiento de voz y una pluralidad de módulos de codificación de vídeo, en donde

la unidad MCU está adaptada para poner en práctica el método de visualización de subtítulos según la reivindicación 1.

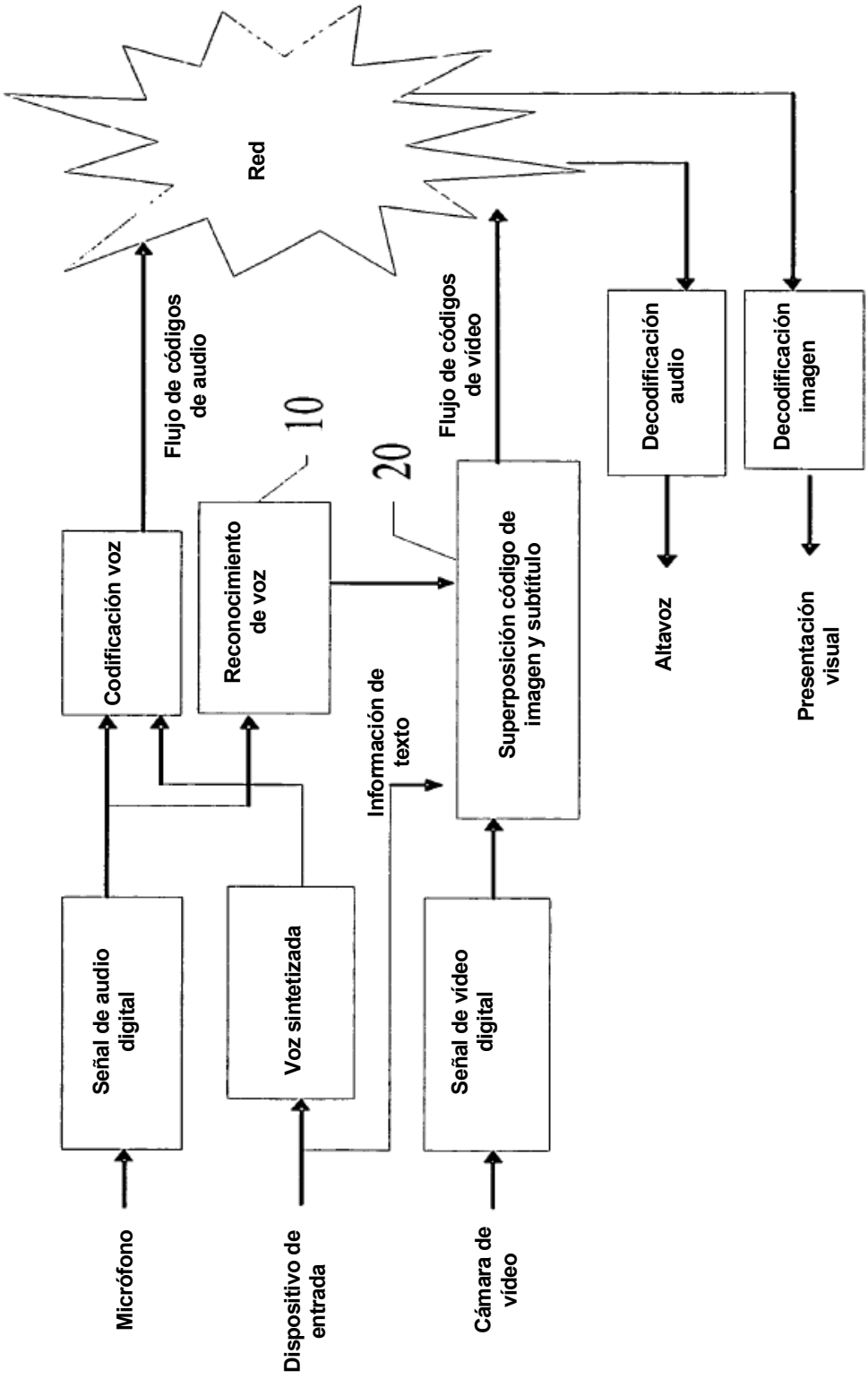
Red
Flujo de códigos de audio
Flujo de códigos de vídeo
10
20
Codificación voz
Reconocimiento de voz
Superposición código de imagen y subtítulo
Decodificación audio
Decodificación imagen
Altavoz
Presentación visual
Información de texto
Señal de audio digital
Voz sintetizada
Señal de vídeo digital
Micrófono
Dispositivo de entrada
Cámara de vídeo

Figura 1

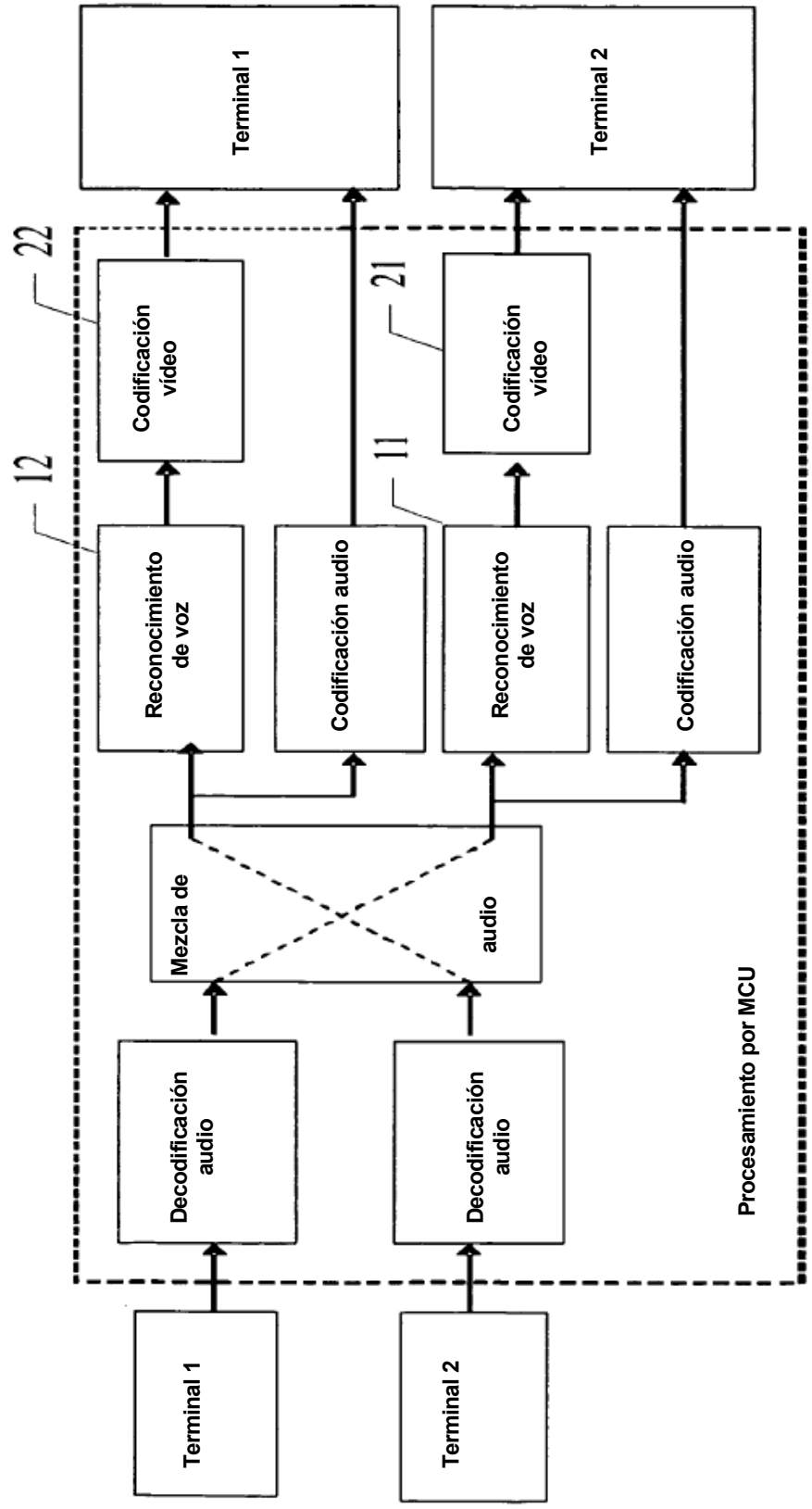
Terminal 1
Terminal 2
22
21
12
11
Codificación vídeo
Codificación vídeo
Reconocimiento de voz
Codificación audio
Reconocimiento de voz
Codificación audio
Mezcla de audio
Decodificación audio
Decodificación audio
Procesamiento por MCU
Terminal 1
Terminal 2

Figura 2

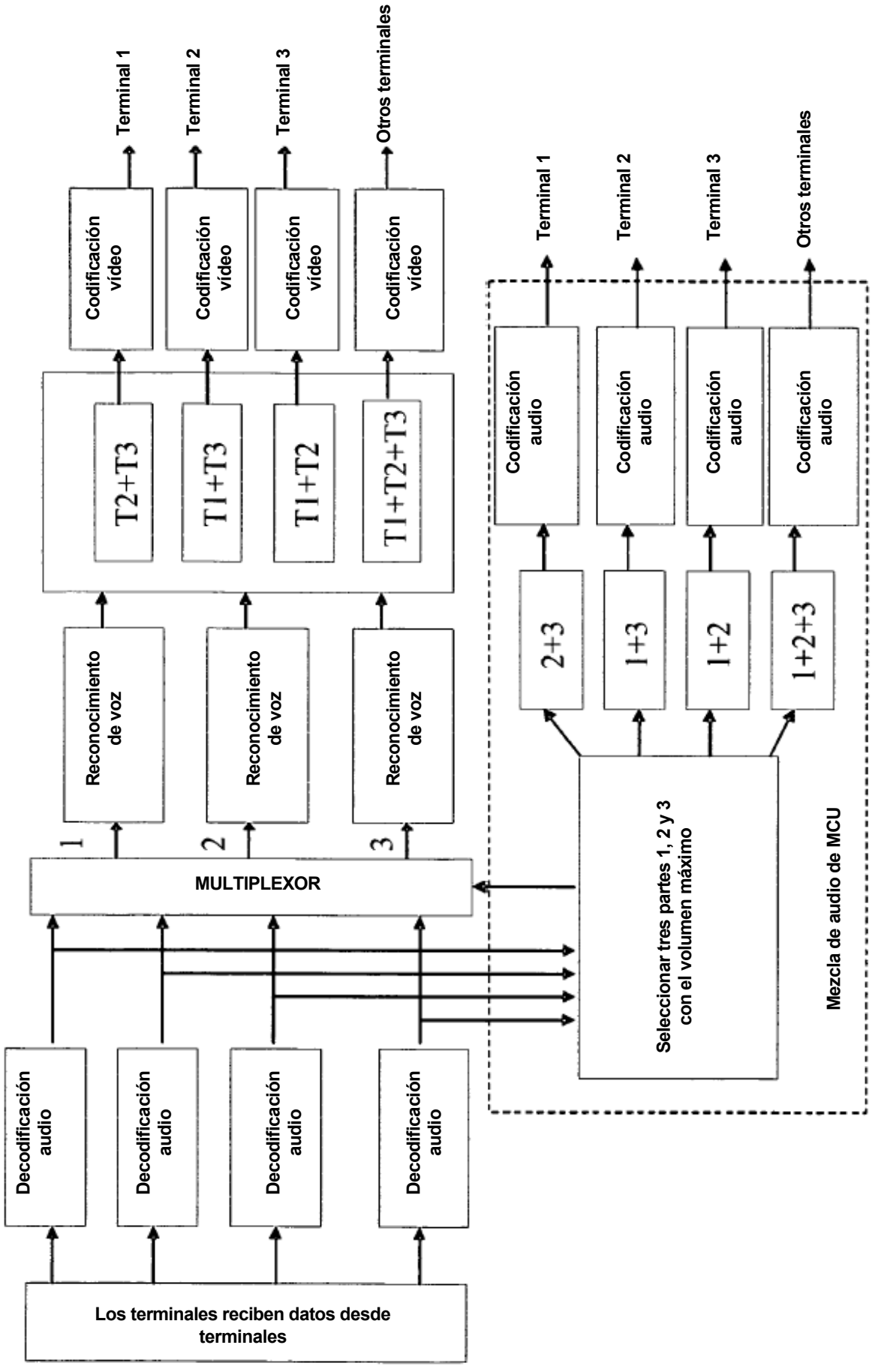
Terminal 1
Terminal 2
Terminal 3
Otros terminales
Codificación vídeo
T2+T3
T1+T3
T1+T2
T1+T2+T3
Reconocimiento de voz
1
2
3
MULTIPLEXOR
Decodificación audio
Los terminales reciben datos desde terminales
Codificación audio
2+3
1+3
1+2
1+2+3
Seleccionar tres partes 1, 2 y 3 con el volumen máximo
Mezcla de audio de MCU

Figura 3

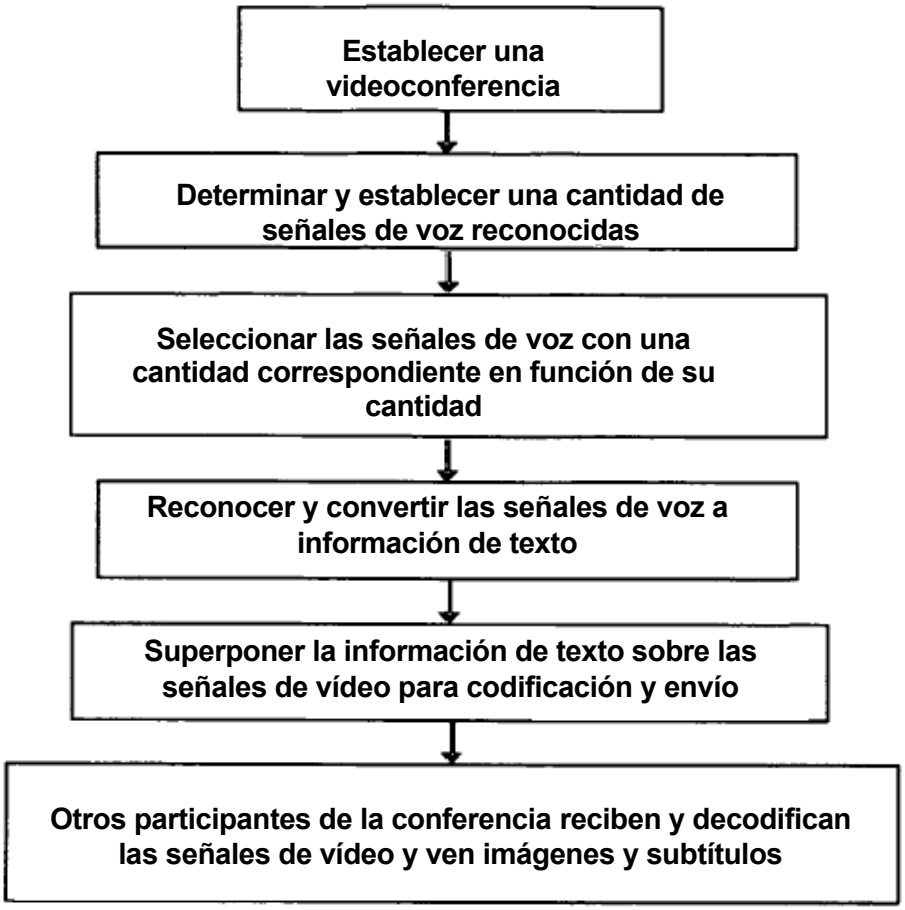
Establecer una videoconferencia
Determinar y establecer una cantidad de señales de voz reconocidas
Seleccionar las señales de voz con una cantidad correspondiente en función de su cantidad
Reconocer y convertir las señales de voz a información de texto
Superponer la información de texto sobre las señales de vídeo para codificación y envío
Otros participantes de la conferencia reciben y decodifican las señales de vídeo y ven imágenes y subtítulos

Figura 4