(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 628 901**

(51) Int. CI.:

G10L 13/033 (2013.01)
G10L 21/003 (2013.01)
G10L 15/00 (2013.01)
G06F 21/31 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **01.02.2013 PCT/US2013/024245**

(87) Fecha y número de publicación internacional: **22.08.2013 WO13122750**

(96) Fecha de presentación y número de la solicitud europea: **01.02.2013 E 13749405 (0)**

(97) Fecha y número de publicación de la concesión europea: **29.03.2017 EP 2815398**

(54) Título: **Prueba de interacción humana de audio basada en la conversión texto-a-voz y la semántica**

(30) Prioridad:

17.02.2012 US 201213399496

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
04.08.2017

(73) Titular/es:

MICROSOFT TECHNOLOGY LICENSING, LLC (100.0%)
One Microsoft Way
Redmond, Washington 98052-6399, US

(72) Inventor/es:

QIAN, YAO;
ZHU, BIN BENJAMIN y
SOONG, FRANK KAO-PING

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Prueba de interacción humana de audio basada en la conversión texto-a-voz y la semántica

Antecedentes

Una Prueba de Interacción Humana (HIP), conocida también como CAPTCHA (Prueba de Turing Pública Completamente Automatizada para Distinguir Ordenadores y Humanos) diferencia un usuario humano con respecto a la programación automática (es decir, un robot).

El objetivo de la mayoría de los esquemas de HIP es evitar un acceso automatizado por parte de un ordenador, al mismo tiempo que permitir el acceso por parte de un humano. Típicamente, este objetivo se afronta proporcionando un método para generar y corregir pruebas que pueden ser superadas fácilmente por la mayoría de las personas, y no por la mayor parte de programas de ordenador.

El documento US 2005/015257 da a conocer una prueba para humanos que incluye el planteamiento de una pregunta que se selecciona para ejercer una capacidad conceptual humana, la obtención de una respuesta a la pregunta, y la comparación de la respuesta con una respuesta correcta que sería proporcionada por un ser humano.

Compendio

Este Sumario se aporta para introducir de manera simplificada una selección de conceptos que se describen de forma adicional posteriormente en la Descripción Detallada. Este Sumario no pretende identificar características clave o características esenciales de la materia en cuestión reivindicada, ni está destinado a usarse con el fin de limitar el alcance de la materia en cuestión reivindicada. Por el contrario, la invención proporciona un proceso implementado por ordenador para proporcionar una prueba de interacción humana automática, y un sistema para generar un desafío basado en audio para una prueba de interacción humana automatizada, según se reivindica posteriormente en la presente.

La técnica de Pruebas de Interacción Humana (HIP) de audio basadas en texto-a-voz descrita en el presente, proporciona una HIP de audio que utiliza tecnologías de texto-a-voz y semántica en la creación de un desafío de audio, con el fin de determinar si un usuario desconocido de un ordenador es un humano o un robot. Con el fin de conseguir que una frase pronunciada sea irreconocible por parte de un sistema genérico o personalizado de Reconocimiento Automático del Habla (ASR) (usado típicamente por robots para intentar descifrar de forma automática una HIP de audio), la técnica evita que el sistema de reconocimiento de habla del sistema de ASR reconozca muestras de HIP generadas por la técnica o que aprenda de ellas. La técnica puede llevar a cabo esto haciendo que las frases de HIP pronunciadas sean muy diferentes con respecto a los datos de audio usados en el entrenamiento de un modelo del sistema de ASR, y variando las características de las palabras o frases de HIP pronunciadas. Típicamente, los sistemas de ASR se basan en modelos estadísticos. Cuanto más lejos esté la frase de HIP con respecto a la distribución de datos de entrenamiento del modelo de ASR, más difícil le resultará al sistema de ASR reconocer la palabra o frase de la HIP.

La técnica de HIP de audio de texto-a-voz descrita en la presente puede aplicar deformación (*warping*) de frecuencia espectral, deformación de la duración de las vocales, adición de ruido de fondo, adición de eco, y espacios de tiempo entre palabras en la generación de una frase de HIP pronunciada, a través de un motor de Texto-a-Voz (TTS), además de la deformación de volumen reivindicada. Por lo tanto, el tempo, la altura tonal y la voz resultantes de la frase pronunciada son muy diferentes con respecto a los datos normales usados para entrenar sistemas de ASR. Además, la técnica utiliza un conjunto grande de parámetros de texto-a-voz para permitirse cambiar efectos de manera frecuente o constante, con el fin de impedir el uso de desafíos de HIP de audio utilizados previamente, para entrenar un modelo de un sistema de ASR con vistas a reconocer desafíos de HIP generados por la técnica.

La técnica de HIP de audio de texto-a-voz descrita en la presente añade un mecanismo adicional para diferenciar usuarios humanos con respecto a robots: para superar el desafío debe entenderse una frase del desafío de HIP de audio. En particular, la frase puede ser una pregunta o una instrucción para la cual se requiere una interpretación semántica de la frase planteada como desafío de audio con el fin de responder correctamente a este último. De esta manera, incluso si el mecanismo previamente descrito falla, es decir, un sistema de ASR puede reconocer todas las palabras de la frase usada como desafío de audio, un robot seguirá sin poder superar la prueba si no entiende la frase. La interpretación semántica de frases se sigue considerando un problema importante de la Inteligencia Artificial.

Descripción de los dibujos

Las características, aspectos y ventajas específicos de la exposición se entenderán mejor en relación con la siguiente descripción, las reivindicaciones adjuntas y los dibujos anexos, en donde:

la FIG. 1 es una arquitectura ejemplificativa para llevar a la práctica una realización ilustrativa de la técnica de HIP de audio de texto-a-voz descrita en la presente.

La FIG. 2 representa un diagrama de flujo de un proceso ejemplificativo para utilizar una realización de la técnica de HIP de audio de texto-a-voz.

La FIG. 3 representa otro diagrama de flujo de un proceso ejemplificativo para utilizar una realización de la técnica de HIP de audio de texto-a-voz.

La FIG. 4 representa todavía otro diagrama de flujo de un proceso ejemplificativo para utilizar una realización de la técnica de HIP de audio de texto-a-voz.

La FIG. 5 es un esquema de un entorno informático ejemplificativo que se puede utilizar para llevar a la práctica la técnica de HIP de audio de texto-a-voz.

Descripción detallada

En la siguiente descripción de la técnica de HIP de audio de texto-a-voz, se hace referencia a los dibujos adjuntos, los cuales forman parte de la misma, y que muestran ejemplos de ilustración por medio de los cuales se puede llevar a la práctica la técnica de HIP de audio de texto-a-voz descrita en la presente. Debe entenderse que pueden utilizarse otras realizaciones y que se pueden realizar cambios estructurales sin desviarse con respecto al alcance de la materia en cuestión reivindicada.

1.0 Técnica de HIP de audio de texto-a-voz

Las siguientes secciones proporcionan una introducción a las pruebas de interacción humana (HIPs), una visión general de la técnica de HIP de audio de texto-a-voz descrita en la presente, así como una arquitectura ejemplificativa y procesos ejemplificativos para llevar a la práctica la técnica. Se proporcionan también detalles de varias realizaciones de la técnica.

1.1 Introducción a las Pruebas de Interacción Humana (HIPs)

Una HIP, conocida también como CAPTCHA (Prueba de Turing Pública Completamente Automatizada para Distinguir Ordenadores y Humanos) diferencia un usuario humano con respecto a la programación automática (es decir, un robot). El objetivo de la mayoría de los esquemas de HIP es evitar un acceso automatizado por parte de un ordenador, al mismo tiempo que permitir el acceso por parte de un humano. Típicamente, este objetivo se afronta proporcionando un método para generar y corregir pruebas que pueden ser superadas fácilmente por la mayoría de las personas, y no por la mayor parte de programas de ordenador.

En la actualidad hay disponibles muchos esquemas de HIP. Por ejemplo, uno de los esquemas visuales convencionales funciona mediante la selección aleatoria de caracteres o palabras de un diccionario, y a continuación la representación de una imagen distorsionada que contiene los caracteres o palabras. Seguidamente, este esquema presenta una prueba a su usuario, el cual consiste en la imagen distorsionada y una solicitud de teclear parte de los caracteres o palabras que aparecen en la imagen. Individualizando los tipos de deformaciones que se aplican, se crea una imagen en la que la mayoría de los humanos pueden leer el número requerido de caracteres o palabras de la imagen distorsionada, mientras que los programas de ordenador actuales típicamente no pueden.

En otro ejemplo de HIP de audio, caracteres individuales son pronunciados por personas. Los caracteres pronunciados que son los mismos que una HIP visual adjunta se distorsionan y se juntan con un espacio de tiempo diferente entre letras. Se adiciona también ruido de fondo. A continuación se le pide a un usuario que teclee las letras pronunciadas.

Todavía en otra HIP de audio, se pronuncian palabras individuales, las mismas se distorsionan y se les añade ruido de fondo. A un usuario se le pide que teclee las palabras pronunciadas. Son tolerables algunos errores en las palabras tecleadas.

1.2 Visión general de la técnica

La técnica de HIP de audio de texto-a-voz descrita en la presente en algunas realizaciones usa frases o palabras diferentes (preferentemente que no se repiten) generadas por medio de un motor de texto-a-voz, en calidad de desafíos de HIP de audio. La técnica puede aplicar diferentes efectos en un sintetizador de texto-a-voz que pronuncia una frase o palabras que se usarán como HIP. Los diferentes efectos pueden incluir, entre otros, por ejemplo, deformación de frecuencia espectral; deformación de duración de las vocales; adición de ruido de fondo; adición de eco; y variación del espacio de tiempo entre palabras, además de la deformación de volumen reivindicada. En algunas realizaciones, la técnica cambia el conjunto de parámetros para generar desafíos de HIP de audio con el tiempo y para la generación de diferentes desafíos, con el fin de evitar que un ASR aprenda un modelo que pueda ser usado para reconocer los desafíos de HIP de audio generados por la técnica. Adicionalmente, en algunas realizaciones, la técnica introduce el requisito de interpretación semántica con el fin de resolver los desafíos de HIP.

1.3 Arquitectura ejemplificativa

La FIG. 1 muestra una arquitectura ejemplificativa 100 para llevar a la práctica una realización de la técnica de HIP de audio de texto-a-voz. Tal como se muestra en la FIG. 1, esta arquitectura ejemplificativa 100 incluye un módulo selector 101 de texto que puede contener una biblioteca 102 de texto (por ejemplo, frases y respuestas esperadas textuales) o una biblioteca 103 de palabras. El módulo 101 selecciona y proporciona texto 106 al módulo 104 de generación de HIP y respuestas esperadas 107 a un verificador 109. En una realización, el módulo selector 101 de texto puede seleccionar un elemento específico (por ejemplo, preferentemente una frase en texto y su respuesta esperada), o bien de forma aleatoria o bien de una manera específica. Las frases de la biblioteca 102 de texto se pueden seleccionar de entre documentos o artículos proporcionados desde una cierta fuente, rastreados en Internet, o generados a partir de un módulo (no mostrado en la FIG. 1) sobre la base de reglas o patrones específicos. En algunas realizaciones, con las frases se producen respuestas esperadas. Una respuesta esperada puede ser la propia frase, o una contestación que proporcionarían seres humanos como respuesta a la frase. La primera se usa típicamente cuando la frase se produce a partir de un documento o artículo de manera automática. La segunda se usa típicamente cuando la frase la produce un módulo de programa.

En una realización, el módulo selector 101 de texto puede contener una biblioteca de palabras de entre las cuales se puede seleccionar un grupo de palabras correlacionadas o sin correlación, o bien aleatoriamente o bien de una cierta manera. Las palabras seleccionadas se usan como texto seleccionado 106 que se envía al módulo 104 de generación de HIP, y las palabras dispuestas en el mismo orden que el texto seleccionado 106 se usan también como respuesta esperada 107 que se envía al verificador 109.

La arquitectura incluye un módulo 104 de generación de HIP de audio que reside en el mismo dispositivo informático general 500 que el selector 101 de texto o uno diferente. El dispositivo informático general 500 se describirá de forma más detallada con respecto a la FIG. 5. El módulo 104 de generación de HIP contiene un motor de TTS, módulos 114 de distorsión de TTS y módulos 116 de distorsión post-TTS. Un motor de TTS convencional consta de dos partes, un modelo 108 de habla y un sintetizador 110 de texto-a-voz. El motor de TTS procesa el texto seleccionado 106 usando el modelo 108 de habla. Entre las dos partes del motor de TTS (el modelo 108 de habla y el sintetizador 110 de TTS), unos módulos 114 de distorsión de TTS ajustan los parámetros según son modelados por el modelo 108 de habla para aplicar una o más distorsiones, de manera que el texto seleccionado 106 se distorsionará cuando sea leído por el sintetizador 110 de texto-a-voz. La salida de TTS puede ser procesada adicionalmente por módulos 116 de distorsión post-TTS para aplicar una o más distorsiones adicionales, tales como un eco o ruido de fondo añadido al texto pronunciado generado por TTS. El sonido resultante se usa como HIP/CATCHA 112 de audio. Las distorsiones o parámetros de distorsión que definen una o más de las distorsiones en el sintetizador 110 de TTS o post-TTS se pueden cambiar aleatoriamente o con un patrón específico, con cada instancia de generación de una cadena de desafío de audio.

El módulo 104 de generación de HIP determina los parámetros de distorsión que se usan para modelar el texto seleccionado usando el modelo 108 de habla. En una realización, este modelo 108 de habla es un Modelo Oculto de Markov (HMM) que se usa para modelar el espectro frecuencial (tracto vocal), la frecuencia fundamental (fuente vocal) y la duración del habla (prosodia). Los módulos 114 de distorsión de TTS en el interior del módulo 104 de generación de HIP pueden incluir un módulo 120 de deformación de frecuencia que deforma los parámetros frecuenciales del texto seleccionado 106 cuando el mismo es leído por el sintetizador 110 de texto-a-voz. Los módulos 114 de distorsión de TTS también pueden incluir un módulo 118 para variar las duraciones de sonidos pronunciables. Por ejemplo, este módulo 118 puede llevar a cabo una deformación de la duración de las vocales, que hace variar el tiempo en el que se pronuncian vocales de la frase seleccionada 106 cuando la misma es leída por el sintetizador 110 de texto-a-voz. Adicionalmente, los módulos 114 de distorsión de TTS incluyen un módulo 122 para variar los volúmenes de sonidos, y pueden incluir un módulo 124 para variar los espacios de tiempo entre palabras cuando el sintetizador 110 de texto-a-voz genera una voz correspondiente al texto seleccionado 106.

Después de que la voz del texto seleccionado sea generada por el sintetizador 110 de texto a voz, pueden aplicarse una o más distorsiones adicionales con los módulos 116 de distorsión post-TTS. Los módulos 116 de distorsión post-TTS pueden incluir un módulo 126 de adición de eco para añadir efectos de eco y/o un módulo 128 de adición de ruido de fondo para añadir sonidos de fondo a un fragmento de audio generado del texto seleccionado 106, desde el sintetizador 110 de texto a voz.

El módulo 128 de adición de ruido de fondo puede añadir diferentes sonidos de fondo. En una realización, como sonidos de fondo puede añadirse música. En otra realización, otra voz, a la que en lo sucesivo se hace referencia como voz de fondo, puede añadirse a la voz, a la que se hace referencia como voz en primer plano, del texto seleccionado 106 desde el sintetizador 110 de texto a voz. A los sonidos de fondo se les pueden aplicar distorsiones y otras modificaciones para producir variaciones adicionales sobre los sonidos de fondo para los mismos desafíos de HIP de audio o desafíos diferentes.

Cuando se añade voz, la voz de fondo puede estar en el mismo idioma que la voz en primer plano. También puede estar en un idioma diferente al idioma de la voz en primer plano. Por ejemplo, cuando la voz en primer plano está en inglés, la voz de fondo puede estar en chino o español. La voz de fondo se puede generar con el sintetizador 110 de TTS, de una manera similar a la voz en primer plano. Durante la generación de la voz de fondo pueden aplicarse

distorsiones diferentes, tales como deformación de frecuencia y otras mencionadas anteriormente para la voz en primer plano. El texto de la voz de fondo puede ser una frase seleccionada de una biblioteca de texto, o palabras seleccionadas aleatoriamente de un diccionario. Con una voz de fondo añadida, los humanos pueden ver fácilmente la diferencia entre los dos idiomas, e identificar y reconocer el idioma en primer plano, pero una máquina, tal como un motor de ASR, no puede diferenciar la voz en primer plano con respecto a la voz de fondo, y, por lo tanto, no puede reconocer el texto pronunciado de la voz en primer plano.

Un desafío de audio generado desde el módulo de generación de HIP se envía a un usuario desconocido 130 el cual puede introducir una respuesta con diversos métodos, tales como usando un teclado, un ratón, o una pantalla táctil. En una realización, el usuario desconocido 130 puede pronunciar una respuesta, y se usan tecnologías de reconocimiento de voz para reconocer la respuesta y convertirla en texto. A continuación, se envía una respuesta 132 de texto recibida al verificador 109 el cual compara la respuesta recibida con la respuesta esperada del desafío de audio. El verificador 109 identifica al usuario desconocido 130 como humano 134, si determina que la respuesta 132 del usuario desconocido coincide con la respuesta esperada 107. En caso contrario, el usuario desconocido se identifica como robot 136. En una realización, el usuario desconocido 130 reconoce un desafío 112 de audio para proporcionar una respuesta correcta con el fin de superar la prueba. En otra realización, el usuario desconocido 130 debe disponer de una interpretación semántica del desafío de audio con el fin de proporcionar una respuesta correcta para superar la prueba.

Se pueden usar muchas técnicas en el verificador 109 para determinar si una respuesta recibida 132 coincide con una respuesta esperada. En una realización, el verificador determina que las dos respuestas coinciden entre sí únicamente cuando las mismas coinciden exactamente. En este caso, no se tolera ningún error. En otra realización, el verificador determina que las dos respuestas coinciden entre sí, si el error entre las dos respuestas está por debajo de un error de tolerancia. En una realización, el error entre dos respuestas se calcula usando la distancia de edición o sus variantes.

El verificador 109 puede procesar una respuesta antes de compararla con la otra respuesta. Por ejemplo, el verificador 109 puede normalizar el texto de una respuesta, tal como sustituyendo una palabra o una cadena de texto con su expresión normalizada, y eliminar palabras irrelevantes. El verificador 109 también puede convertir una respuesta de texto en una cadena de fonos, y comparar cadenas de fonos para determinar si las dos respuestas coinciden o no entre sí. Se pueden usar muchas técnicas para convertir texto en fonos. En una realización, el modelo de habla en un TTS se usa para convertir texto en fonos.

1.4 Procesos ejemplificativos para llevar a la práctica la técnica

En general, la FIG. 2 muestra un proceso ejemplificativo general para llevar a la práctica una realización de la técnica de HIP de audio de texto-a-voz descrita en la presente. Tal como se muestra en el bloque 202, de entre una pluralidad de frases de texto o una pluralidad de palabras se selecciona una frase de texto o un grupo de palabras correlacionadas o sin correlación. Tal como se muestra en el bloque 204, se aplica un motor de texto-a-voz para generar la voz del texto seleccionado como desafío de audio, con vistas a la identificación de si un usuario desconocido es un humano o un robot, con una o más distorsiones aplicadas durante o tras la generación de la voz del texto seleccionado.

La FIG. 3 muestra un proceso ejemplificativo 300 más detallado para llevar a la práctica otra realización de la técnica de HIP de audio de texto-a-voz. En general, esta realización de la técnica de HIP de audio de texto-a-voz funciona en primer lugar encontrando o definiendo una biblioteca de frases de texto o palabras discretas, tal como se muestra en el bloque 302. Por ejemplo, las frases de texto se pueden seleccionar de entre varias fuentes de texto apropiadas. En una realización, el texto de la biblioteca de texto es frases no repetidas que se extraen automáticamente de artículos o documentos cuya interpretación es sencilla para humanos, tales como periódicos y revistas que estaban destinados a lectores humanos comunes. En una realización, la longitud de las frases extraídas está en un intervalo preestablecido. Si una frase es demasiado corta, se puede combinar con la siguiente frase o simplemente se puede descartar. Una frase demasiado larga se puede dividir en dos o más trozos más pequeños para que encajen con la longitud requerida. Los artículos o documentos se pueden proporcionar desde fuentes internas o se pueden rastrear en Internet. En otra realización, la técnica de HIP de audio construye o define una biblioteca de palabras. Esta biblioteca se puede construir a partir de un diccionario eliminando palabras irrelevantes o palabras confusas. Las palabras que pueden provocar fácilmente confusión en los humanos por su deletreo o sus sonidos se pueden eliminar de la biblioteca. Dada esta biblioteca de frases de texto o palabras, tal como se muestra en el bloque 304, la técnica selecciona automáticamente una frase de texto de la biblioteca de frases de texto, o selecciona un grupo de palabras correlacionadas o sin correlación, de entre la biblioteca de palabras, para su uso en la creación de un desafío de audio con el fin de determinar si un usuario desconocido es un humano o un robot. La técnica también puede recuperar la respuesta esperada a partir de la biblioteca de las frases de texto, en caso de que la respuesta esté almacenada en la biblioteca con la frase de texto recuperada, o genera una respuesta esperada a partir de las frases de texto o grupo de palabras correlacionadas o sin correlación seleccionadas, tal como se muestra en el bloque 306. En una realización, la respuesta esperada generada es la misma que la cadena seleccionada de texto. En otra realización, la respuesta esperada generada es el resultado después de que se aplique la normalización de texto al texto seleccionado. La normalización de texto recibe una entrada de texto y produce una salida de texto que convierte el texto introducido a un formato normalizado. Por ejemplo, palabras

irrelevantes tales como “un”, “una” pueden eliminarse, “I’m” (en inglés) se puede sustituir por “I am” (en inglés) durante la normalización del texto. (Tal como se describirá posteriormente, la respuesta esperada se envía a un verificador 316 para su comparación con la respuesta de un usuario desconocido 314, con el fin de determinar si el usuario desconocido 318 es un humano o un robot 320). A continuación, el texto seleccionado se procesa automáticamente para determinar parámetros con el fin de añadir una o más distorsiones cuando el texto seleccionado sea leído por un sintetizador de texto-a-voz, tal como se muestra en el bloque 308. En la determinación de los parámetros pueden usarse uno o más modelos de lenguaje en el bloque 308. Estas distorsiones, que se describen de forma más detallada posteriormente, pueden incluir deformación de frecuencia espectral, deformación de duración de las vocales, deformación de espacios de tiempo entre palabras, además de la deformación de volumen reivindicada.

Una vez que el sintetizador de texto-a-voz produce la voz del texto seleccionado en el bloque 308, la técnica crea un desafío de audio en el bloque 310. Pueden aplicarse una o más distorsiones adicionales a la voz generada en el bloque 308 durante la creación de un desafío de audio. Estas distorsiones pueden ser la adición de eco, voz de fondo o música. Pueden aplicarse distorsiones a la música o voz de fondo antes de añadirla a la voz generada en el bloque 308. La voz de fondo se puede generar de una manera similar a la generación de la voz en primer plano, por ejemplo, seleccionando una frase de texto o un grupo de palabras correlacionadas o sin correlación, de entre una biblioteca, y, a continuación, aplicando un modelo de lenguaje y un sintetizador de texto-a-voz para generar la voz de fondo. Pueden determinarse parámetros y los mismos se pueden modificar para aplicar una o más distorsiones cuando la voz es generada por el sintetizador de texto-a-voz. Estas distorsiones pueden ser similares a las distorsiones aplicadas dentro del sintetizador de TTS durante la generación de la voz en primer plano. La voz de fondo puede estar en diferentes idiomas. En una realización, la voz de fondo añadida está en el mismo idioma que el correspondiente a la voz en primer plano generada en el bloque 308. En otra realización, la voz de fondo añadida está en un idioma diferente al idioma de la voz en primer plano generada en el bloque 308. La adición de distorsiones durante y después de la generación de la voz usando el sintetizador de TTS sirve para crear un desafío de audio cuyo reconocimiento es relativamente sencillo para una persona, pero difícil para un ordenador, e introduce variaciones entre desafíos de audio generados.

Una vez que en el bloque 310 se genera un desafío de audio, la siguiente etapa es enviar y presentar el desafío de audio a un usuario desconocido para su identificación, tal como se muestra en el bloque 312. A continuación, se solicita al usuario desconocido que responda tecleando o pronunciando una respuesta al desafío de audio, tal como se muestra en el bloque 314. Debe indicarse que, incluso cuando la respuesta esperada es la cadena de texto seleccionada, un atacante no puede tomar el desafío de HIP de audio como respuesta pronunciada, puesto que el reconocimiento del habla no puede convertir correctamente la respuesta pronunciada en una respuesta de texto que se usa en el siguiente bloque. Tal como se muestra en el bloque 316, la respuesta de este usuario se compara a continuación con la respuesta esperada. En una realización, se pronuncia la respuesta del usuario. Se aplican técnicas de reconocimiento de voz para convertir la respuesta pronunciada en una respuesta de texto antes de la comparación con la respuesta esperada. Únicamente si se determina que la respuesta tecleada coincide con la respuesta esperada, se considera que el usuario desconocido es un humano (bloque 318). Si no, se considera que el usuario desconocido es un robot (bloque 320). En una realización, se requiere que la coincidencia sea exacta. En otra realización, la coincidencia no tiene que ser exacta. Puede permitirse alguna disparidad entre las dos respuestas. Se sigue determinando que la respuesta del usuario coincide con la respuesta esperada, siempre que la disparidad se sitúe por debajo de alguna tolerancia o umbral de error predeterminado.

En la determinación de si la respuesta del usuario coincide con la respuesta esperada, el verificador en el bloque 316 puede normalizar las respuestas para eliminar variantes de las mismas expresiones antes de comparar las dos respuestas. Esta normalización puede eliminar caracteres o palabras irrelevantes, y sustituir una o más palabras con palabras equivalentes, normalizadas. Por ejemplo, “I’m” (en inglés) se puede sustituir por “I am” (en inglés), e “intl.” se puede sustituir por internacional. Todavía en otra realización, las respuestas se pueden convertir en cadenas de sonidos, es decir, fonos, y la comparación se basa en los fonos en lugar del texto.

En el bloque 316 se pueden usar muchas técnicas para calcular errores entre dos respuestas. En una realización, se usa la distancia de adición para calcular los errores entre dos cadenas de texto o fonos. La fase de normalización mencionada en el párrafo anterior se puede aplicar antes de calcular la distancia de edición. El cálculo de la distancia de edición se puede basar en palabras o fonos, o en caracteres. Cuando el error se calcula basándose en palabras, dos palabras pueden considerarse iguales si una de ellas es una variante de la otra, tal como la forma plural de la otra palabra, o la diferencia entre las dos palabras está dentro de cierto margen de tolerancia de error. Cuando el error se calcula basándose en fonos, dos fonos pronunciados de manera similar se pueden considerar iguales en el cálculo de errores de dos respuestas.

La FIG. 4 muestra todavía otro proceso ejemplificativo 400 para llevar a la práctica otra realización de la técnica de HIP de audio de texto-a-voz. En general, en esta realización, la técnica funciona en primer lugar definiendo una biblioteca de frases de texto discretas y sus respuestas esperadas que requieren que un usuario interprete el significado semántico de la frase, con el fin de proporcionar una respuesta correcta a la misma, tal como se muestra en el bloque 402. En una realización, el texto es frases no repetidas que se generan automáticamente basándose en un conjunto preestablecido de reglas. Una frase de texto de la biblioteca es típicamente una instrucción o una

pregunta para la cual se requiere una interpretación de la frase con el fin de proporcionar una respuesta correcta. Por ejemplo, un conjunto de reglas puede producir muchas preguntas en relación con la suma o resta de elementos, donde un elemento puede ser cualquier objeto común tal como, por ejemplo, una manzana, un perro o un avión. Usando diferentes números y elementos, pueden generarse muchas preguntas, tales como “Simón comió tres manzanas ayer y se ha comido dos plátanos hoy. ¿Cuál es el número total de frutas que ha comido Simón desde ayer?” El tema, el tiempo, los números y los nombres de los elementos se pueden cambiar para generar más preguntas. Otro conjunto de reglas puede producir muchas preguntas usando la multiplicación y/o la división, y/o la suma y la multiplicación. Como ejemplo alternativo, un conjunto de reglas puede generar una pregunta pidiendo a un usuario que introduzca una respuesta de una manera específica, tal como proporcionando una frase y, a continuación, pidiendo a un usuario que introduzca la segunda letra de las palabras pronunciadas en el orden inverso, o que introduzca la tercera palabra seguida por la anterior a la misma. Este conjunto de reglas también puede generar muchas preguntas. La técnica de HIP de audio de texto-a-voz hace que varíe el patrón de las frases generadas por el mismo conjunto de reglas, e intercala frases generadas usando diferentes conjuntos de reglas con el fin de generar desafíos de HIP de audio para evitar que robots clasifiquen correctamente una HIP de audio, basándose en el conjunto de reglas usadas en la generación de los desafíos de HIP de audio, o mediante conocimiento de cómo proporcionar una respuesta correcta basándose en ciertos patrones o palabras clave.

Las frases de texto se almacenan con sus contestaciones apropiadas o respuestas esperadas. Dada esta biblioteca de frases de texto, tal como se muestra en el bloque 404, la técnica selecciona automáticamente de la biblioteca una o más frases de texto, para su uso en la creación de un desafío de audio que se usará en la determinación de si un usuario de ordenador desconocido es un humano o un robot. A continuación, esta frase seleccionada se puede procesar automáticamente para determinar una o más distorsiones que se pueden añadir cuando la primera sea leída por un sintetizador de texto-a-voz, tal como se muestra en el bloque 406. Estas distorsiones, que se describen de forma más detallada posteriormente, incluyen deformación de frecuencia espectral, deformación de duración de vocales, y variaciones del espacio de tiempo entre palabras, además de la deformación de volumen reivindicada. A la voz generada por el sintetizador de texto-a-voz se le pueden aplicar una o más distorsiones adicionales, tales como adición de ruido de fondo y adición de eco, en la creación de una HIP de audio, tal como se muestra en el bloque 408. No obstante, debe indicarse que, en una realización, la frase que requiere interpretación semántica no se distorsiona cuando la misma es leída por el sintetizador de texto-a-voz o después de ello. El desafío de HIP de audio sin distorsión se basa en una interpretación semántica del desafío para determinar si un usuario desconocido es un humano o un robot. La interpretación semántica evita que robots proporcionen una respuesta correcta.

La siguiente etapa, tal como se muestra en el bloque 410, consiste en presentar el desafío de audio a la parte desconocida, con vistas a su identificación. A continuación, se solicita a la parte desconocida que responda a la frase que requiere interpretación semántica, o bien tecleando o bien pronunciando una respuesta apropiada, según se muestra en el bloque 412. Una respuesta pronunciada se puede convertir en una respuesta de texto aplicando técnica de reconocimiento de voz. Una respuesta se puede convertir en una cadena de fonos que representan cómo se pronuncian la respuesta. Se puede aplicar normalización en una respuesta para sustituir variantes con formas normalizadas de expresar la respuesta, y también se pueden eliminar caracteres o palabras irrelevantes. A continuación, la respuesta del usuario se compara con la respuesta esperada del desafío de audio para determinar si las mismas coinciden o no, tal como se muestra en el bloque 414. Únicamente si se determina que la respuesta del usuario coincide con la respuesta esperada, se considera que el usuario desconocido es un humano, tal como se muestra en el bloque 416. En caso contrario, se considera que el usuario desconocido es un robot, tal como se muestra en el bloque 418. En una realización, se determina que dos respuestas coinciden entre sí únicamente si las mismas coinciden entre ellas exactamente. En otra realización, se determina que dos respuestas coinciden entre sí si su error está dentro de un margen de tolerancia. Pueden usarse técnicas diferentes para calcular el error de las respuestas, por ejemplo, la distancia de edición y sus variantes. Muchas de las técnicas descritas para el proceso ejemplificativo mostrado en la FIG. 3 también se pueden aplicar para el proceso ejemplificativo descrito en la FIG. 4.

1.5 Detalles de varias realizaciones de la técnica

Tras haberse descrito una arquitectura ejemplificativa y procesos ejemplificativos para llevar a la práctica la técnica de HIP de audio de texto-a-voz, los siguientes párrafos proporcionan diversos detalles para implementar varias realizaciones de la técnica.

1.5.1 Pueden aplicarse diversas distorsiones

Tal como se ha descrito anteriormente, pueden aplicarse una o más distorsiones durante la creación de un desafío de audio a partir del texto seleccionado. Estas distorsiones se pueden aplicar cuando se genera la voz del texto y/o después de ello. Las distorsiones se pueden cambiar con cada instancia de producción de un desafío de audio. La técnica de HIP de audio de texto-a-voz puede utilizar deformación de frecuencia espectral, variación de sonidos pronunciables, tal como deformación de la duración de vocales, y variaciones del espacio de tiempo entre palabras vecinas cuando el texto seleccionado es leído por un sintetizador de texto-a-voz, y/o mediante la adición de ruido de fondo y adición de eco en la voz generada en la creación de un desafío de HIP de audio, además de las variaciones reivindicadas de los volúmenes de la voz. A continuación, se describen detalles de la aplicación de estas y otras distorsiones en la creación de un desafío de HIP de audio usado para determinar si un usuario desconocido es un humano o un robot.

1.5.1.1 Deformación de frecuencia espectral

Pueden aplicarse muchos tipos diferentes de deformación de frecuencia cuando el texto seleccionado se convierte en voz para distorsionar la voz generada, con el fin de conseguir que el reconocimiento del desafío de audio sea más difícil para un robot. Por ejemplo, pueden aplicarse una o más distorsiones por deformación de frecuencia durante la generación del desafío de audio para distorsionar la voz generada. Para llevar a cabo esto, se determinan varias funciones y parámetros de deformación, y los mismos se usan con el fin de variar efectos de deformación de frecuencia espectral a lo largo del tiempo y a lo largo de diferentes desafíos de audio.

En una realización de la técnica de HIP de audio de texto-a-voz, para llevar a cabo la deformación de frecuencia espectral, se usa una función de deformación con un parámetro α, y α puede cambiar a lo largo del tiempo t. Al mismo tiempo, se usa una función $\widehat{\omega} = \psi_{\alpha(t)}(\omega)$ para llevar a cabo la transformación. Las funciones de deformación pueden ser lineales, lineales por tramos, bilineales o no lineales. En una realización, la técnica de audio de texto-a-voz descrita en la presente usa una función bilineal de deformación de frecuencia basada en un simple filtro de primer orden pasa-todo con ganancia unidad,

$$\psi_{\alpha(t)}(z) = \frac{z^{-1} - \alpha(t)}{1 - \alpha(t) z^{-1}},$$

o

$$Y_{a(t)}(w) = w + 2\tan^{-1}\frac{a(t)\text{sen}(w)}{1 - a(t)\cos(w)},$$

donde $|\alpha(t)| < 1$.

En una realización, el parámetro de deformación $\alpha(t)$ preferentemente cambia progresivamente a lo largo del tiempo. Así, en este caso se usa una función sinusoidal tal como la siguiente:

$$a(t) = B + A \text{ sen } ((k + t)/T * 2 * p)$$

donde A, B y T son el intervalo deformación, el centro de la deformación y el periodo de la deformación, y se fijan o bien manualmente o bien varían dentro de ciertos intervalos, y donde k es la fase inicial y se fija a un valor dentro de [0, T, -1], de manera o bien aleatoria o bien no aleatoria.

Debe indicarse que la función de deformación antes descrita es una función de deformación ejemplificativa que se puede utilizar con la técnica descrita en el presente documento. Pueden usarse otras diversas funciones de deformación, y estas otras funciones de deformación o sus parámetros también pueden variar a lo largo del tiempo o se pueden aplicar progresivamente con el paso del tiempo.

1.5.1.2 Deformación de duración de las vocales

En una realización de la técnica de HIP de audio de texto-a-voz, se varía la duración de la pronunciación de sonidos pronunciables para distorsionar la voz generada de la cadena de texto seleccionada, cuando el texto es leído por un sintetizador de texto-a-voz. Por ejemplo, en una realización, se usa la deformación de la duración de vocales para variar las duraciones de pronunciaciones de vocales en la lectura del texto seleccionado por parte del sintetizador de texto-a-voz. En esta realización que utiliza deformación de duración de vocales, la técnica de HIP de audio de texto-a-voz en primer lugar fija una duración mínima y máxima para cada vocal la cual puede seguir siendo percibida por personas, y, a continuación, ajusta aleatoriamente la duración de las vocales durante la generación de la voz del texto seleccionado, por parte del sintetizador de texto-a-voz. Debe indicarse también que ciertas consonantes también pueden variarse de una manera similar

1.5.1.3 Deformación de volumen

La deformación de volumen se aplica para cambiar los volúmenes de sonidos pronunciables cuando el texto seleccionado es leído por el sintetizador de texto-a-voz. En particular, se fijan un volumen mínimo y un volumen máximo, y se aplica a una pronunciación un volumen aleatorio entre los volúmenes mínimo y máximo para aplicar la deformación de volumen.

1.5.1.4 Variación del espacio de tiempo entre palabras

El espacio de tiempo entre dos palabras también se puede variar cuando el texto seleccionado es leído por el sintetizador de texto-a-voz. En una realización, se fijan un espacio de tiempo mínimo y un espacio de tiempo máximo, y se puede seleccionar aleatoriamente un espacio de tiempo entre el espacio de tiempo mínimo y el espacio de tiempo máximo, y el mismo se puede aplicar al espacio de tiempo de dos palabras vecinas. Si el espacio

de tiempo seleccionado es negativo, las dos palabras vecinas se pronuncian con un solapamiento especificado. Esta variación de espacios de tiempo entre palabras puede hacer que resulte complicado para un sistema de ASR segmentar una frase en palabras individuales.

1.5.1.5 Adición de ruido de fondo y de eco

La técnica de HIP de audio de texto-a-voz también puede añadir una o más distorsiones a la voz generada del texto seleccionado. En algunas realizaciones, se pueden aplicar ruido de fondo y eco a la voz leída por el sintetizador de texto-a-voz. Por ejemplo, el ruido de fondo puede ser simple ruido, música, voz hablada en el mismo idioma u otro, y otras opciones del estilo. A la voz generada del texto seleccionado también se le puede añadir eco. Por ejemplo, se pueden fijar aleatoriamente el porcentaje de disminución, el tiempo de retardo y el volumen de eco inicial. Adicionalmente, una o más distorsiones aplicadas después de la generación de la voz de la cadena de texto seleccionada pueden incluir la adición de otra voz generada mediante una técnica de texto-a-voz, al ruido de fondo de la voz de la cadena de texto para crear un desafío de audio. En una realización, esta voz adicional añadida al ruido de fondo puede estar en un idioma diferente al de la cadena de texto seleccionada. El habla de fondo se puede seleccionar de manera que sea un idioma no conocido por la mayoría de personas seleccionadas como objetivo del desafío de audio generado. Las personas pueden identificar fácilmente las hablas de diferentes idiomas y concentrarse en el habla en primer plano que es conocida por el usuario humano. Los robots pueden encontrar dificultades en discernir el habla en primer plano con respecto al habla de fondo, y, por lo tanto, no pueden reconocer el habla en primer plano. En otra realización, el habla de fondo puede estar en el mismo idioma que el habla en primer plano. El habla de fondo se puede generar leyendo una frase o un grupo de palabras correlacionadas o sin correlación, con un sintetizador de texto-a-voz. El volumen del habla de fondo se puede variar en un intervalo apropiado para conseguir que el habla en primer plano sea fácilmente identificada por humanos. Pueden aplicarse una o más distorsiones cuando se añade el ruido de fondo. Por ejemplo, al habla de fondo añadida se le pueden aplicar una o más distorsiones cuando dicha habla de fondo es leída por un sintetizador de texto-a-voz o después de ello. Estas distorsiones pueden incluir, aunque sin carácter limitativo, deformación de frecuencia, deformación de la duración de sonidos pronunciables, deformación de volumen, y variaciones de espacios de tiempo entre palabras. Al habla de fondo generada por un sintetizador de texto-a-voz se le pueden aplicar una o más distorsiones. Por ejemplo, puede añadirse eco al habla de fondo antes de que la misma sea añadida al habla en primer plano. Además, el habla de fondo puede presentarse en forma de un habla sin sentido o audio grabado. En la realización en la que el habla de fondo está en el mismo idioma que el habla de fondo, el habla de fondo sin sentido puede ayudar a las personas a identificar y reconocer el habla en primer plano.

1.5.2 Texto usado en un desafío de HIP de audio

En algunas realizaciones de la técnica de HIP de audio de texto-a-voz, cada desafío de HIP de audio es una frase pronunciada a través de un sintetizador de texto-a-voz. Una realización simple de la técnica de HIP de audio de texto-a-voz selecciona aleatoriamente una frase con la longitud apropiada de palabras, típicamente dentro de un intervalo específico, a partir de un artículo, y usa el sintetizador de texto-a-voz para expresar la frase seleccionada. En otras realizaciones, un desafío de HIP de audio es una cadena de palabras correlacionadas o sin correlación pronunciadas a través de un sintetizador de texto-a-voz. Estas palabras se pueden seleccionar de entre una biblioteca de palabras, se pueden construir a partir de un diccionario eliminando palabras que pueden crear confusión en las personas cuando estas últimas reconocen estas palabras, y palabras irrelevantes.

La técnica presenta un desafío de audio a un usuario desconocido y le pide al usuario desconocido que teclee o pronuncie una respuesta para el desafío de audio. En algunas realizaciones, al usuario desconocido se le pide que responda con la frase o la cadena de palabras que ha oído. Esto se usa típicamente cuando no es necesaria la interpretación semántica del texto seleccionado. Solamente se necesita que el usuario desconocido reconozca correctamente la frase o cadena de palabras pronunciada. Estas realizaciones presentan la ventaja de que pueden generar fácilmente desafíos de HIP de audio de diferentes idiomas. En otras realizaciones, es necesario que el usuario desconocido entienda la frase pronunciada para proporcionar una respuesta correcta. Esta frase es típicamente una instrucción o una pregunta generada automáticamente con uno o más conjuntos de reglas. Estas realizaciones presentan la ventaja de que, en los desafíos de audio generados, se aplica un nivel de seguridad adicional. Es necesario que el usuario desconocido no solamente reconozca correctamente la frase pronunciada, sino que también entienda correctamente la frase para proporcionar una respuesta correcta. Cuando se requiere la interpretación semántica para responder a un desafío de audio, la contestación esperada se genera típicamente con la frase, y se almacena junto con la misma en una biblioteca.

1.5.3 Interpretación semántica

Aunque muchos de los desafíos anteriores de HIP de audio generados por medio de la técnica no requieren una interpretación semántica de las frases usadas como desafío de audio, en realizaciones de la técnica de HIP de audio de texto-a-voz, se añade un mecanismo adicional para ayudar a diferenciar humanos de robots. En particular, para superar la prueba se requiere una interpretación de la frase de un desafío de HIP de audio. Esta frase puede ser una pregunta o una instrucción. Por ejemplo, en algunas realizaciones la técnica define una pluralidad de categorías de preguntas o instrucciones basándose en los tipos de una pregunta o instrucción. Pueden asociarse una o más reglas a cada categoría para ayudar a generar frases de texto y sus contestaciones esperadas automáticamente. Se

requiere la interpretación semántica de dicha frase para proporcionar una respuesta correcta. Las personas entienden la frase, y, por lo tanto, pueden proporcionar fácilmente una contestación correcta. Por otro lado, los robots no poseen la capacidad de entender la frase, y, por lo tanto, no pueden proporcionar una contestación correcta. Por ello, la propia frase es un desafío de HIP. Si la frase se usa como texto seleccionado para generar un desafío de audio, incluso si los robots reconocen correctamente el texto del desafío de audio, siguen sin poder proporcionar una respuesta correcta y superar la prueba de HIP, puesto que no entienden el significado semántico de la frase. Al sistema se le pueden añadir tipos adicionales de preguntas e instrucciones. En una realización, una de las categorías es que una respuesta esperada sea una cadena específica de caracteres o palabras basadas en una frase. Por ejemplo, puede ser una frase aleatoriamente seleccionada, seguida por una instrucción para pedir a un usuario que introduzca la segunda letra de las palabras de la frase previa, o que introduzca las últimas dos palabras en el orden inverso, etcétera. El conjunto de reglas asociadas a la categoría determina tipos diferentes de instrucciones (y, por lo tanto, respuestas esperadas diferentes para la misma frase seleccionada) y maneras diferentes de expresar instrucciones equivalentes que producen la misma respuesta esperada. Puesto que los robots no entienden la instrucción, no serían capaces de proporcionar una respuesta correcta. Una vez que se ha generado una frase compuesta del tipo mencionado (una frase seleccionada aleatoriamente más la instrucción sucesiva), se genera también la respuesta esperada. La respuesta o respuestas esperadas se pueden añadir a una biblioteca que se seleccionará posteriormente en la generación de un desafío de HIP de audio. En otra realización, una categoría a utilizar es que una contestación esperada sea un resultado de cálculo específico. Por ejemplo, el conjunto de reglas asociadas a la categoría es la generación de diferentes preguntas relacionadas con resultados de cálculos y diferentes formas de expresiones que producen el mismo resultado de cálculo. Por ejemplo, una frase generada puede ser: “Simón se comió tres manzanas ayer y se ha comido dos plátanos hoy, ¿Qué día se comió más frutas en términos de unidades de fruta?” La contestación esperada a esta frase se genera también automáticamente. Variando el tema, el tiempo, la pregunta a realizar, y formas equivalentes de expresar una misma cosa, la técnica puede generar una pluralidad de frases y sus respuestas esperadas.

2.0 Entornos operativos ejemplificativos:

La técnica de HIP de audio de texto-a-voz descrita en la presente se puede hacer funcionar dentro de numerosos tipos de entornos o configuraciones de sistemas informáticos de propósito general o propósito especial. La FIG. 5 ilustra un ejemplo simplificado de un sistema de ordenador de propósito general en el cual se pueden implementar varias realizaciones y elementos de la técnica de HIP de audio de texto-a-voz, que se ha descrito en la presente. Debe indicarse que cualquier recuadro que se represente mediante líneas discontinuas o de trazos en la FIG. 5 representa realizaciones alternativas del dispositivo informático simplificado, y que cualquiera o la totalidad de estas realizaciones alternativas, que se describen posteriormente, se pueden usar en combinación con otras realizaciones alternativas que se describen durante todo este documento.

Por ejemplo, la FIG. 5 muestra un diagrama de sistema general que presenta un dispositivo informático simplificado 500. Dichos dispositivos informáticos se pueden encontrar típicamente en dispositivos que tienen por lo menos cierta capacidad computacional mínima, incluyendo, aunque sin carácter limitativo, ordenadores personales, ordenadores servidores, dispositivos informáticos de mano, ordenadores portátiles o móviles, dispositivos de comunicaciones tales como teléfonos celulares y PDAs, sistemas de multiprocesador, sistemas basados en microprocesadores, unidades de adaptación de televisores, electrónica de consumo programable, PCs en redes, miniordenadores, ordenadores centrales, reproductores de medios de audio o vídeo, etcétera.

Para permitir que un dispositivo implemente la técnica de HIP de audio de texto-a-voz, el dispositivo debe tener una capacidad computacional y una memoria de sistemas suficientes para posibilitar operaciones computacionales básicas. En particular, tal como se ilustra mediante la FIG. 5, la capacidad computacional se ilustra generalmente mediante una o más unidad(es) 510 de procesado, y también puede incluir una o más GPUs 515, estando cualquiera de ellas o ambas en comunicación con la memoria 520 del sistema. Obsérvese que la(s) unidad(es) 510 de procesado del dispositivo informático general puede(n) ser microprocesadores especializados, tales como un DSP, un VLIW, u otro micro-controlador, o puede(n) ser CPUs convencionales que tengan uno o más núcleos de procesado, incluyendo núcleos basados en GPU especializados en una CPU multi-núcleo.

Además, el dispositivo informático simplificado de la FIG. 5 también puede incluir otros componentes, tales como, por ejemplo, una interfaz 530 de comunicaciones. El dispositivo informático simplificado de la FIG. 5 también puede incluir uno o más dispositivos convencionales 540 de entrada de ordenador (por ejemplo, dispositivos señaladores, teclados, dispositivos de entrada de audio, dispositivos de entrada de vídeo, dispositivos de entrada hápticos, dispositivos para recibir transmisiones de datos por cable o inalámbricas, etcétera). El dispositivo informático simplificado de la FIG. 5 también puede incluir otros componentes opcionales, tales como, por ejemplo, uno o más dispositivos convencionales 550 de salida de ordenador (por ejemplo, dispositivo(s) 555 de visualización, dispositivos de salida de audio, dispositivos de salida de vídeo, dispositivos para comunicar transmisiones de datos por cable o inalámbricas, etcétera). Obsérvese que las interfaces 530 de comunicaciones, los dispositivos 540 de entrada, los dispositivos 550 de salida, y los dispositivos 550 de almacenamiento típicos, para ordenadores de propósito general, son bien conocidos para aquellos versados en la materia, y no se describirán de forma detallada en la presente.

El dispositivo informático simplificado de la FIG. 5 también puede incluir una variedad de soportes legibles por

ordenador. Los soportes legibles por ordenador pueden ser cualquier soporte disponible al que se pueda acceder mediante ordenador 500 a través de dispositivos 560 de almacenamiento, e incluye soportes tanto volátiles como no volátiles que sean o bien extraíbles 570 ó bien/y no extraíbles 580, para el almacenamiento de información, tal como instrucciones legibles o ejecutables por ordenador, estructuras de datos, módulos de programa, u otros datos. A título de ejemplo, y sin carácter limitativo, los soportes legibles por ordenador pueden comprender soportes de almacenamiento de ordenador y soportes de comunicación. Los soportes de almacenamiento de ordenador incluyen, aunque sin carácter limitativo, soportes legibles por ordenador o máquina o dispositivos de almacenamiento, tales como DVDs, CDs, discos flexibles, unidades de cinta, unidades de disco duro, unidades ópticas, dispositivos de memoria de estado sólido, RAM, ROM, EEPROM, memoria *flash* u otra tecnología de memorias, casetes magnéticos, cintas magnéticas, medios de almacenamiento de disco magnético, u otros dispositivos de almacenamiento magnético, o cualquier otro dispositivo que se pueda usar para almacenar la información deseada y al que se puede acceder mediante uno o más dispositivos informáticos.

El almacenamiento de información, tal como instrucciones legibles o ejecutables por ordenador, estructuras de datos, módulos de programa, etcétera, también se puede lograr usando cualquiera de una variedad de los soportes de comunicación antes mencionados, para codificar una o más señales de datos moduladas u ondas portadoras, u otros mecanismos de transporte o protocolos de comunicaciones, e incluye cualquier mecanismo de distribución de información por cable o inalámbrico. Obsérvese que las expresiones “señal de datos modulada” u “onda portadora” se refieren en general a una señal en la que se han fijado o cambiado una más de sus características de tal manera que codifican información en la señal. Por ejemplo, los soportes de comunicación incluyen soportes por cable, tales como una conexión de red por cable o de cableado directo que transporta una o más señales de datos moduladas, y soportes inalámbricos, tales como soportes acústicos, de RF, de infrarrojos, de láser y otros inalámbricos para transmitir y/o recibir una o más señales de datos moduladas u ondas portadoras. Las combinaciones de cualesquiera de las anteriores también deben incluirse dentro del alcance de los soportes de comunicación.

Además, software, programas, y/o productos de programa de ordenador que materialicen parte o la totalidad de las diversas realizaciones de la técnica de HIP de audio de texto-a-voz descrita en la presente, o partes de la misma, se pueden almacenar, recibir, transmitir o leer desde cualquier combinación deseada de soportes legibles por ordenador o máquina o dispositivos de almacenamiento, y soportes de comunicación en forma de instrucciones ejecutables por ordenador u otras estructuras de datos.

Finalmente, la técnica de HIP de audio de texto-a-voz descrita en la presente se puede describir además en el contexto general de instrucciones ejecutables por ordenador, tales como módulos de programa, que sean ejecutados por un dispositivo informático. En general, los módulos de programa incluyen rutinas, programas, objetos, componentes, estructuras de datos, etcétera, que llevan a cabo tareas particulares o implementan tipos de datos abstractos particulares. Las realizaciones descritas en la presente también se pueden llevar a la práctica en entornos informáticos distribuidos, en donde las tareas son ejecutadas por uno o más dispositivos de procesado remotos, o dentro de una nube de uno o más dispositivos, que se enlazan a través de una o más redes de comunicaciones. En un entorno informático distribuido, los módulos de programa pueden estar situados en soportes de almacenamiento de ordenador tanto locales como remotos, incluyendo dispositivos de almacenamiento de medios. Todavía adicionalmente, las instrucciones antes mencionadas se pueden implementar, de forma parcial o en su totalidad, como circuitos lógicos de hardware, los cuales pueden incluir o no un procesador.

Debe indicarse también que cualquiera o la totalidad de las realizaciones alternativas antes mencionadas y que se describen en la presente, se pueden usar en cualquier combinación deseada para formar realizaciones híbridas adicionales. Aunque la materia objeto de la invención se ha descrito en un lenguaje específico de características estructurales y/o acciones metodológicas, debe entenderse que la materia objeto definida en las reivindicaciones adjuntas no se limita necesariamente a las características o acciones específicas descritas anteriormente. Las características y acciones específicas antes descritas se dan a conocer como formas ejemplificativas de implementación de las reivindicaciones.

REIVINDICACIONES

1. Proceso implementado por ordenador para proporcionar una prueba de interacción humana automática, que comprende:

seleccionar (202; 304; 404) una cadena de texto de entre una pluralidad de frases de texto o una pluralidad de palabras, comprendiendo la cadena de texto seleccionada una pregunta o instrucción referente al texto;

aplicar (204; 310; 408) un motor de texto-a-voz a la cadena de texto seleccionada para generar un desafío de audio, requiriendo el desafío de audio un conocimiento semántico de la pregunta o instrucción a responder, comprendiendo además la aplicación de una o más distorsiones durante la generación del desafío de audio mediante el cambio de los volúmenes de sonidos pronunciables de la cadena de texto seleccionada con el fin de crear el desafío de audio, y comprendiendo la aplicación de un volumen aleatorio entre un volumen mínimo y un volumen máximo a cada uno de los sonidos pronunciables;

recibir (314; 412) una respuesta al desafío de audio desde un usuario desconocido; y

verificar (316; 414) la respuesta al desafío de audio del usuario desconocido, para determinar si el usuario desconocido es un humano o un robot.

2. Proceso implementado por ordenador de la reivindicación 1, en donde la respuesta recibida es pronunciada por el usuario desconocido, y en donde se aplica reconocimiento de voz para reconocer la respuesta y comparar la respuesta con una contestación correcta.

3. Proceso implementado por ordenador de la reivindicación 1, en donde la aplicación de la o las distorsiones comprende además aplicar deformación de frecuencia espectral.

4. Proceso implementado por ordenador de la reivindicación 1, en donde la aplicación de la o las distorsiones comprende además ajustar espacios de tiempo entre palabras pronunciadas.

5. Proceso implementado por ordenador de la reivindicación 1, que comprende además:

distorsionar (308; 406) los parámetros de un modelo de habla de la cadena de texto seleccionada con una o más distorsiones, de manera que la cadena de texto seleccionada se distorsionará cuando sea leída por el motor de texto-a-voz;

usando los parámetros distorsionados y el modelo de habla, leer la cadena de texto seleccionada para generar el desafío de audio con el uso de un sintetizador de texto-a-voz; y

determinar automáticamente (316; 414) si una respuesta del usuario desconocido coincide con una respuesta esperada.

6. Sistema para generar un desafío basado en audio, para una prueba de interacción humana automatizada, que comprende:

un dispositivo informático (500) de propósito general;

un programa de ordenador que comprende módulos de programa ejecutables por el dispositivo informático de propósito general, en donde el dispositivo informático es dirigido por los módulos de programa del programa de ordenador para llevar a cabo el proceso de cualquier reivindicación anterior.

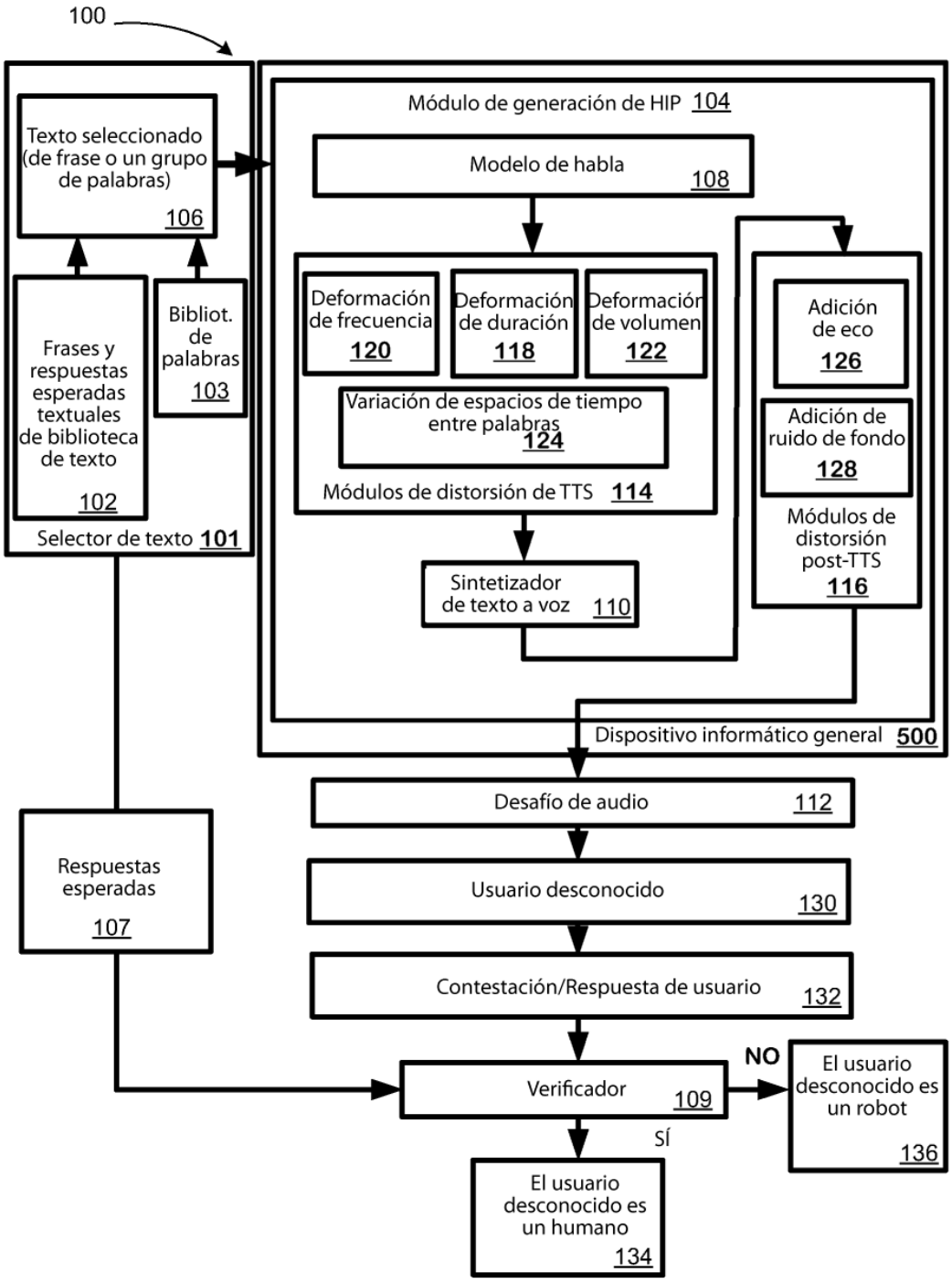
100
Texto seleccionado (de frase o un grupo de palabras) 106
Frases y respuestas esperadas textuales de biblioteca de texto 102
Bibliot. de palabras 103
Selector de texto 101
Módulo de generación de HIP 104
Modelo de habla 108
Deformación de frecuencia 120
Deformación de duración 118
Deformación de volumen 122
Variación de espacios de tiempo entre palabras 124
Módulos de distorsión de TTS 114
Sintetizador de texto a voz 110
Adición de eco 126
Adición de ruido de fondo 128
Módulos de distorsión post-TTS 116
Dispositivo informático general 500
Desafío de audio 112
Usuario desconocido 130
Contestación/Respuesta de usuario 132
Respuestas esperadas 107
Verificador 109
NO
El usuario desconocido es un robot 136
SÍ
El usuario desconocido es un humano 134

FIG. 1

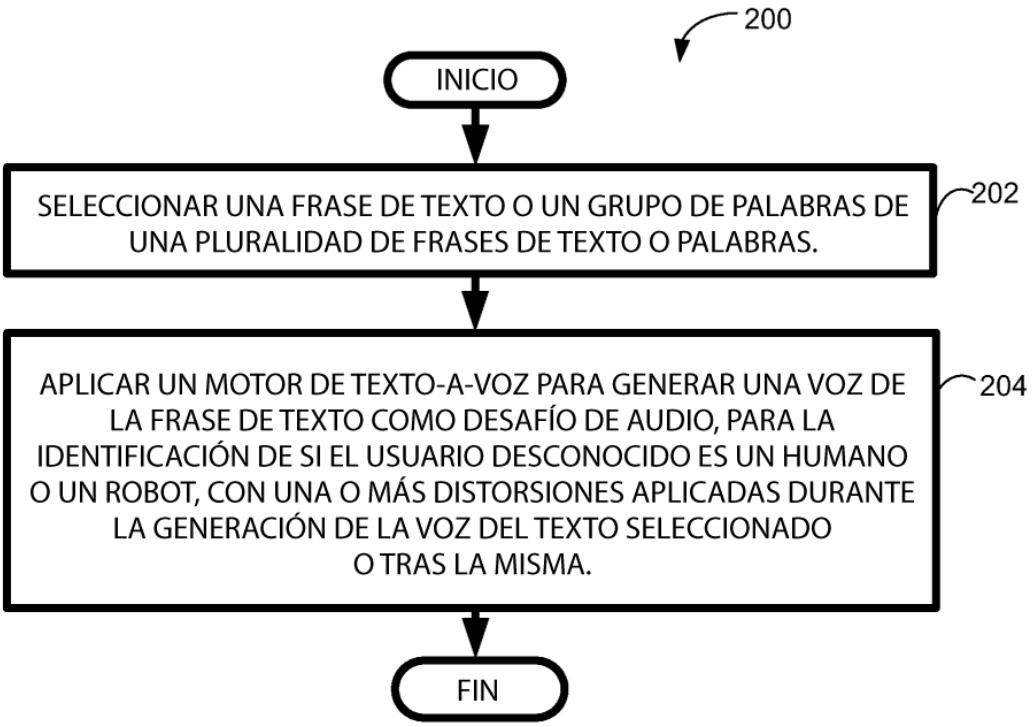
200
INICIO
SELECCIONAR UNA FRASE DE TEXTO O UN GRUPO DE PALABRAS DE UNA PLURALIDAD DE FRASES DE TEXTO O PALABRAS.
202
APLICAR UN MOTOR DE TEXTO-A-VOZ PARA GENERAR UNA VOZ DE LA FRASE DE TEXTO COMO DESAFÍO DE AUDIO, PARA LA IDENTIFICACIÓN DE SI EL USUARIO DESCONOCIDO ES UN HUMANO O UN ROBOT, CON UNA O MÁS DISTORSIONES APLICADAS DURANTE LA GENERACIÓN DE LA VOZ DEL TEXTO SELECCIONADO O TRAS LA MISMA.
204
FIN

FIG. 2

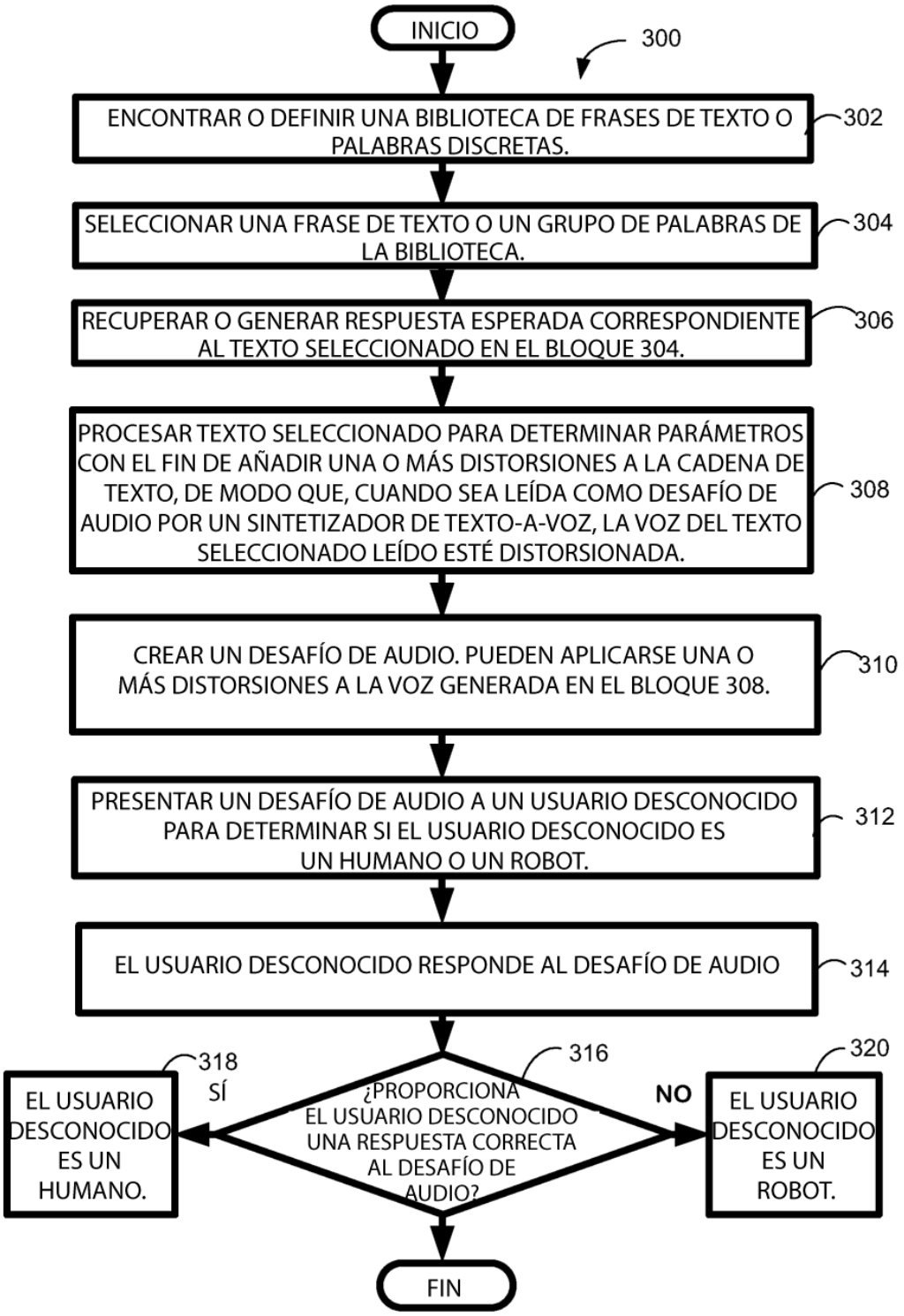
INICIO
300
ENCONTRAR O DEFINIR UNA BIBLIOTECA DE FRASES DE TEXTO O PALABRAS DISCRETAS.
302
SELECCIONAR UNA FRASE DE TEXTO O UN GRUPO DE PALABRAS DE LA BIBLIOTECA.
304
RECUPERAR O GENERAR RESPUESTA ESPERADA CORRESPONDIENTE AL TEXTO SELECCIONADO EN EL BLOQUE 304.
306
PROCESAR TEXTO SELECCIONADO PARA DETERMINAR PARÁMETROS CON EL FIN DE AÑADIR UNA O MÁS DISTORSIONES A LA CADENA DE TEXTO, DE MODO QUE, CUANDO SEA LEÍDA COMO DESAFÍO DE AUDIO POR UN SINTETIZADOR DE TEXTO-A-VOZ, LA VOZ DEL TEXTO SELECCIONADO LEÍDO ESTÉ DISTORSIONADA.
308
CREAR UN DESAFÍO DE AUDIO. PUEDEN APLICARSE UNA O MÁS DISTORSIONES A LA VOZ GENERADA EN EL BLOQUE 308.
310
PRESENTAR UN DESAFÍO DE AUDIO A UN USUARIO DESCONOCIDO PARA DETERMINAR SI EL USUARIO DESCONOCIDO ES UN HUMANO O UN ROBOT.
312
EL USUARIO DESCONOCIDO RESPONDE AL DESAFÍO DE AUDIO
314
316
¿PROPORCIONA EL USUARIO DESCONOCIDO UNA RESPUESTA CORRECTA AL DESAFÍO DE AUDIO?
SÍ
318
EL USUARIO DESCONOCIDO ES UN HUMANO.
NO
320
EL USUARIO DESCONOCIDO ES UN ROBOT.
FIN

FIG. 3

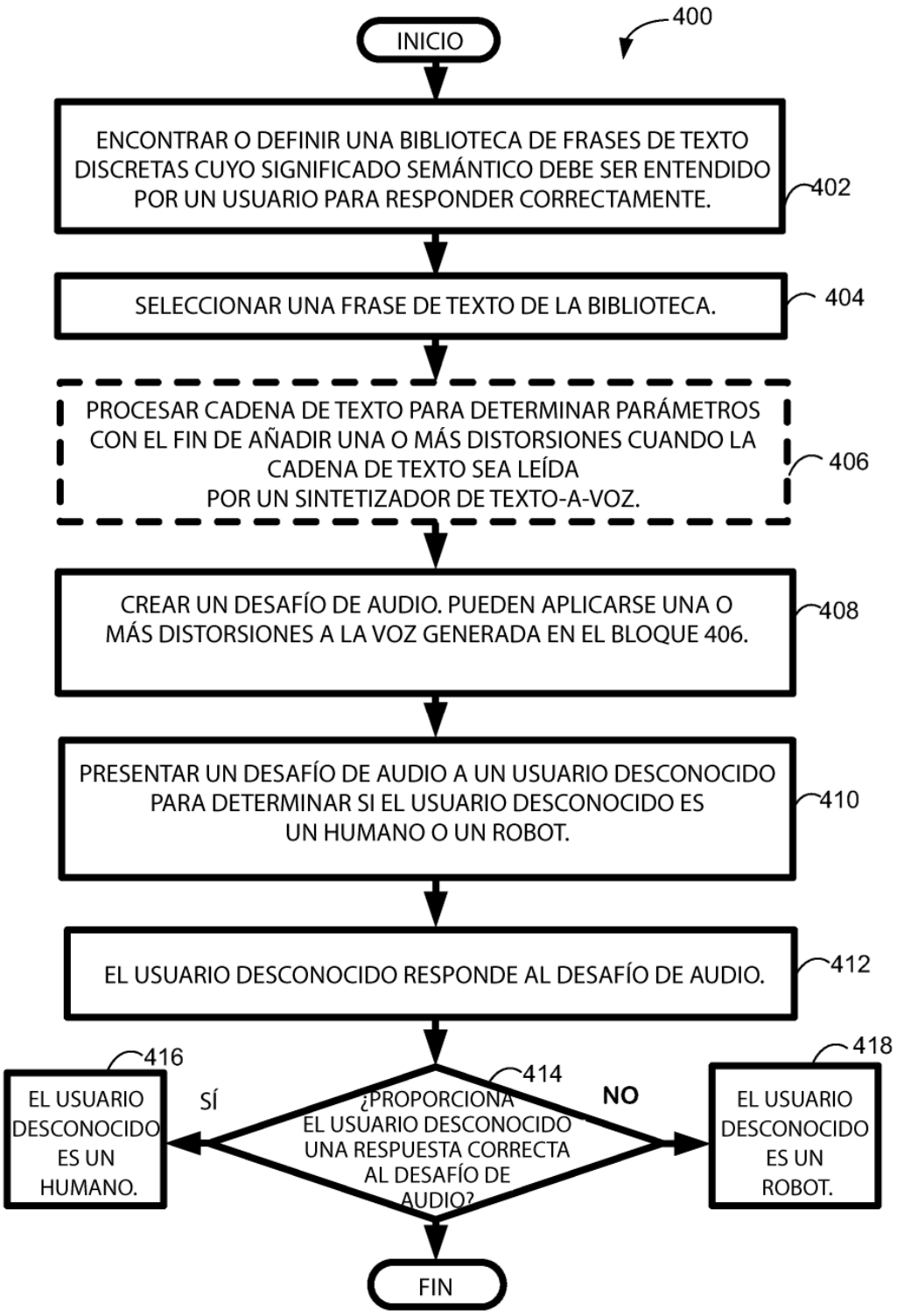
400
INICIO
ENCONTRAR O DEFINIR UNA BIBLIOTECA DE FRASES DE TEXTO DISCRETAS CUYO SIGNIFICADO SEMÁNTICO DEBE SER ENTENDIDO POR UN USUARIO PARA RESPONDER CORRECTAMENTE.
402
SELECCIONAR UNA FRASE DE TEXTO DE LA BIBLIOTECA.
404
PROCESAR CADENA DE TEXTO PARA DETERMINAR PARÁMETROS CON EL FIN DE AÑADIR UNA O MÁS DISTORSIONES CUANDO LA CADENA DE TEXTO SEA LEÍDA POR UN SINTETIZADOR DE TEXTO-A-VOZ.
406
CREAR UN DESAFÍO DE AUDIO. PUEDEN APLICARSE UNA O MÁS DISTORSIONES A LA VOZ GENERADA EN EL BLOQUE 406.
408
PRESENTAR UN DESAFÍO DE AUDIO A UN USUARIO DESCONOCIDO PARA DETERMINAR SI EL USUARIO DESCONOCIDO ES UN HUMANO O UN ROBOT.
410
EL USUARIO DESCONOCIDO RESPONDE AL DESAFÍO DE AUDIO.
412
¿PROPORCIONA EL USUARIO DESCONOCIDO UNA RESPUESTA CORRECTA AL DESAFÍO DE AUDIO?
414
SÍ
NO
EL USUARIO DESCONOCIDO ES UN HUMANO.
416
EL USUARIO DESCONOCIDO ES UN ROBOT.
418
FIN

FIG. 4

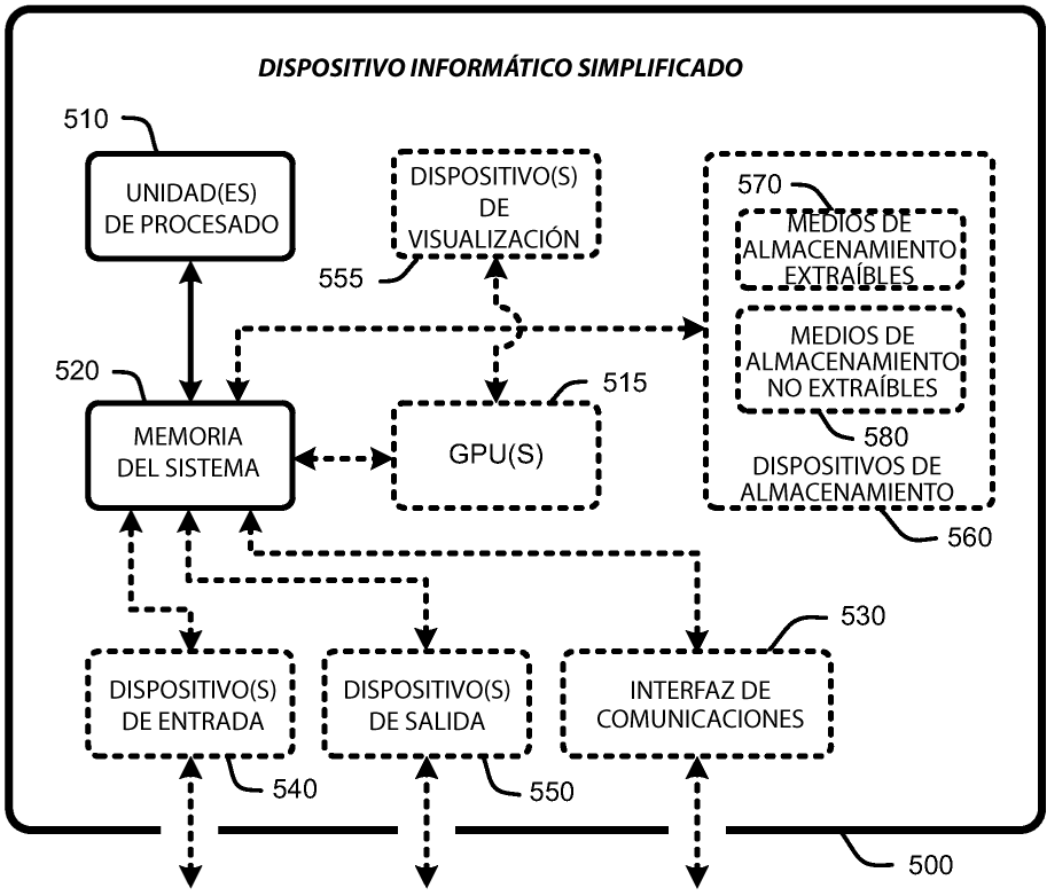
DISPOSITIVO INFORMÁTICO SIMPLIFICADO
510
UNIDAD(ES) DE PROCESADO
DISPOSITIVO(S) DE VISUALIZACIÓN
555
570
MEDIOS DE ALMACENAMIENTO EXTRAÍBLES
MEDIOS DE ALMACENAMIENTO NO EXTRAÍBLES
580
DISPOSITIVOS DE ALMACENAMIENTO
560
520
MEMORIA DEL SISTEMA
515
GPU(S)
530
DISPOSITIVO(S) DE ENTRADA
DISPOSITIVO(S) DE SALIDA
INTERFAZ DE COMUNICACIONES
540
550
500

FIG. 5