(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 577 705**

(51) Int. CI.:

G10L 15/26 (2006.01)
H04M 3/42 (2006.01)
G10L 15/08 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **04.02.2013 E 13382034 (0)**

(97) Fecha y número de publicación de la concesión europea: **06.04.2016 EP 2763136**

(54) Título: **Procedimiento y sistema para obtener información relevante de una comunicación por voz**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
18.07.2016

(73) Titular/es:

TELEFÓNICA, S.A. (100.0%)
Gran Vía, 28
28013 Madrid, ES

(72) Inventor/es:

URDIALES DELGADO, DIEGO

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Procedimiento y sistema para obtener información relevante de una comunicación por voz

Campo técnico de la invención

La presente invención se refiere, en general, al campo del reconocimiento de voz y, más específicamente, se refiere al análisis de comunicaciones por voz para extraer automáticamente información relevante de dicha comunicación por voz y convertirla potencialmente en texto.

Antecedentes de la invención

El "reconocimiento de voz" es la traducción de palabras dichas en texto. También se conoce como "reconocimiento de voz automático" (ASR). En la actualidad, los sistemas de reconocimiento de voz automático (ASR) van dirigidos principalmente al análisis de flujos de audio individuales, tales como aquellos que proceden de un dictado de voz, una orden de voz o un mensaje de buzón de voz. Ejemplos de tales sistemas incluyen aquellos que se usan en aplicaciones comerciales de transcripción de correo de voz, en aplicaciones de asistente virtual, etc. Aunque sería útil tener también una representación textual del contenido de las llamadas de voz, los sistemas de ASR actuales apenas se dirigen al análisis de flujos de audio interrelacionados tales como aquellos que proceden de una llamada de voz.

La investigación en el reconocimiento de voz se ha caracterizado por la acumulación constante de pequeñas mejoras incrementales. También ha habido una tendencia a concentrarse en tareas más difíciles debido tanto al progreso en el rendimiento del reconocimiento de voz como a la disponibilidad de ordenadores más rápidos. La transcripción de voz a texto automatizada para mensajes cortos de voz, tales como aquellos que se dejan en los sistemas de buzón de voz, o para el dictado de SMS u órdenes, ha estado disponible como un servicio con altos niveles de precisión durante algún tiempo. El siguiente paso natural sería extender esta funcionalidad a llamadas de voz, es decir, comunicaciones por voz más largas en las que dos o más partes se comunican de manera síncrona a través de voz. Sin embargo, resulta evidente que una transcripción completa, palabra por palabra, de una llamada de voz no es la manera óptima de transmitir el significado de la conversación de voz para un usuario.

El problema de analizar el contenido de una llamada de voz a través del reconocimiento de voz automático tiene varias particularidades que lo diferencian de otras tareas relacionadas con el ASR:

En primer lugar, como al menos dos partes están incluidas en la comunicación por voz, hay al menos dos flujos de audio para una única comunicación. Según el sistema en el que se realice la comunicación por voz y la configuración de la arquitectura, puede alimentarse con un único flujo de audio el sistema de ASR, que contiene el audio agregado de todas las partes de la comunicación por voz; o pueden alimentarse flujos de audio separados, uno para cada participante individual de la comunicación por voz. Esto difiere de otras tareas relacionadas con el ASR, en las que hay un único flujo de audio no agregado alimentando al motor del ASR.

Acústicamente, como el audio de la llamada de voz se transmite por flujo entre las partes en la llamada a través de una red, está sujeto a interferencias, pérdidas de paquetes y otras degradaciones. Estas degradaciones no se producen cuando el audio se graba y envía a través de la red, en lugar de transmitirse por flujo en tiempo real.

Lingüísticamente, el audio de la llamada de voz es bastante diferente a otros tipos de audio de voz. Como dos o más partes están hablando en la llamada, cada canal de audio contiene habitualmente largos periodos en silencio, correspondientes a tiempos en los que ese orador está escuchando a otra parte. Las llamadas de voz son una forma de comunicación en línea (la sesión es en vivo y la voz se transmite a medida que se pronuncia), es más probable que contengan vacilaciones, frases construidas de manera inapropiada, palabras o frases incompletas, sonidos no verbales y palabras onomatopéyicas. En general, las llamadas de voz contienen una menor proporción de palabras lingüísticamente coherentes que otras formas de comunicación. Y, como las partes en una llamada de voz habitualmente se conocen entre sí y la naturaleza en tiempo real de una llamada de voz permite aclarar inmediatamente cualquier malentendido, las llamadas de voz incluyen elipsis, referencias a conceptos o palabras no mencionados y, en general, expresiones del lenguaje que sólo conocen las partes participantes, con mayor frecuencia que otros tipos de audio de voz.

La técnica anterior incluye ejemplos tales como la solicitud de patente estadounidense US 2008/0201143, que describe un sistema y un procedimiento para extraer el contenido de conversaciones de audio, centrando su atención en (pero sin limitarse a) la monitorización de llamadas realizadas por presos en instituciones correccionales. El sistema descrito en ese documento extrae determinadas características del audio de la llamada de voz. Sin embargo, esa solicitud de patente es fundamentalmente un sistema de etiquetado o control, cuyo objetivo es identificar fragmentos irregulares, sospechosos o poco habituales dentro de la conversación de voz.

Otra solución de la técnica anterior se proporciona en los sistemas y procedimientos de transcripción inteligente de llamadas divulgados en la solicitud estadounidense US2010/158213. Una transcripción de una llamada telefónica es creada y suplementada con información adicional, automáticamente o a petición. Puede añadirse información adicional cuando se detectan palabras clave, tal como agregar expansión de acrónimos cuando se detecta un

acrónimo, o agregar información de identificación a una tarea importante cuando se detecta la mención de la tarea. Por lo tanto, se centra principalmente en el enriquecimiento de una transcripción de audio incluyendo información adicional.

La patente estadounidense US 7599475 B2 propone un procedimiento y un aparato para revelar aspectos comerciales u organizativos de una organización en señales de audio capturadas desde interacciones, difusiones u otros orígenes. El procedimiento y el aparato activan un proceso para detectar sucesos dentro de la señal de audio, y activan luego un proceso adicional, que consume más recursos, alrededor de los sucesos detectados, realzando así la eficacia del proceso. Los sucesos detectados y la salida del proceso adicional son analizados por un proceso de análisis para revelar aspectos comerciales, términos u otros sucesos en la señal de audio.

Además la solicitud de patente estadounidense US 2007/0071206, centrada en la monitorización de conversaciones en instituciones correccionales, contempla tanto la separación de llamadas de voz en canales (diarización) como el aprovechamiento de la información de prosodia para mejorar la detección de sucesos poco habituales en llamadas de voz, o localizar aquellas llamadas que podrían requerir una revisión manual. Sin embargo, está limitada porque no incluye un enfoque sistemático para el uso de diferentes aspectos de la información de audio, y no sólo prosodia y texto, de manera individual o en combinación. De nuevo, esto sigue el objetivo de controlar o etiquetar llamadas, no de extraer la información relevante contenida en la llamada de voz.

Cuando se aplican a flujos de audio interrelacionados tales como aquellos que proceden de llamadas de voz, los sistemas de ASR actuales, como los mencionados anteriormente, tienen diversas desventajas. Debido a la naturaleza lingüística del audio de la llamada de voz, en el que muchas de las palabras no tienen un significado lingüístico, muchas de las frases están incompletas y también abundan las palabras que son onomatopéyicas o están incompletas, los sistemas que pretenden proporcionar una transcripción completa para la llamada de voz (motores de transcripción) proporcionan demasiada información que es de poca utilidad.

Para superar esta desventaja, pueden usarse los sistemas de ASR que pretenden devolver sólo palabras o frases coherentes que se dicen en la llamada de voz, tales como el sistema de ASR de localización de palabras clave.

Sumario de la invención

La presente invención soluciona el problema, mencionado anteriormente, de analizar múltiples flujos de audio de voz de una llamada de voz para obtener palabras significativas que contienen la información relevante de dicha llamada. Por tanto, de acuerdo a la reivindicación 1, se presenta un procedimiento para obtener información relevante de una comunicación por voz proporcionada, entre al menos dos usuarios, en el que la comunicación por voz comprende al menos dos flujos de voz (21) pertenecientes a dichos al menos dos usuarios, comprendiendo el procedimiento las siguientes etapas:

a) generar marcadores con sellos temporales, asociados a dichos al menos dos flujos de voz según determinadas características extraídas de dichos al menos dos flujos de voz (21), en donde, cuando cierta característica es común (55) para ambos flujos de voz, generar un marcador adicional con sello temporal (54) que indica una correlación entre dichos al menos dos flujos de voz (21);

b) extraer (57) determinadas piezas de información de dichos al menos dos flujos de voz según cierta información que comprenden los marcadores;

c) transcribir (58) las determinadas piezas de información en texto para obtener la información relevante de la comunicación por voz.

Los canales pueden no estar disponibles por separado y el procedimiento puede incluir una etapa adicional para solucionarlo, que consiste en separar dichos al menos dos flujos de voz, pertenecientes a dichos al menos dos usuarios, de la comunicación por voz.

En una realización de la invención, las características extraídas de los flujos de voz se toman de la siguiente lista: entonación, volumen, velocidad del habla y ruido. A partir de estas características se procesan los flujos de voz para generar los marcadores con sellos temporales. Dichos marcadores con sellos temporales contienen cierta información, que puede organizarse en una tupla que comprende elementos de la siguiente lista: un identificador para el flujo de voz, un sello temporal de inicio, un sello temporal de finalización, una indicación del tipo de marcador y otros detalles del marcador.

La invención propuesta puede comprender la etapa adicional de comparar una característica de los flujos de voz con un patrón configurado previamente. Por tanto, los marcadores generados pueden comprender información adicional tal como un nivel de confianza, o la indicación de cuáles de los patrones configurados previamente coincidieron.

Una realización de la invención también usa la transcripción de voz a texto y la localización de palabras para generar marcadores. Por tanto, los marcadores generados pueden comprender el nivel de confianza de la transcripción o las palabras localizadas.

Una realización de la invención comprende extraer la pieza de información contenida dentro de los sellos temporales de inicio y finalización definidos en el marcador. Según el tipo de marcador, la acción adoptada puede ser extraer la pieza de información, del mismo flujo de voz, o de otro flujo de voz distinto al indicado por el identificador de la tupla, contenida después del sello temporal de inicio definido en el marcador para una duración específica.

La invención propuesta puede comprender pasar las piezas de información extraída a través de un sistema de reconocimiento de voz automático, basado en la gramática, con una gramática específica. Y, optativamente, las piezas de información extraída pueden hacerse pasar a través de un sistema de reconocimiento de voz automático con un modelo de lenguaje específico.

Como etapa adicional, el procedimiento puede comprender almacenar la comunicación por voz en un dispositivo de almacenamiento de datos, aunque en algunas realizaciones todas las etapas se realizan en tiempo real.

Otro aspecto de la invención, según lo reivindicado en la reivindicación 2, se refiere a un sistema para obtener información relevante de una comunicación por voz proporcionada entre al menos dos usuarios, en donde la comunicación de voz comprende al menos dos flujos de voz pertenecientes a dichos al menos dos usuarios, comprendiendo el sistema:

- un generador de marcadores (22) que recibe dichos al menos dos flujos de voz (21) para ser analizados y generar marcadores con sellos temporales, de acuerdo a ciertas características extraídas de dichos al menos dos flujos de voz, en donde, cuando cierta característica es común (55) a ambos flujos de voz, generar un marcador adicional con sello temporal (54) que indica una correlación entre dichos al menos dos flujos de voz;

- un procesador de acciones (24) para extraer determinadas piezas de información de dichos al menos dos flujos de voz, de acuerdo a alguna información comprendida por los marcadores

y transcribir las determinadas piezas de información en texto, obteniendo así la información relevante de la comunicación por voz.

Los flujos pueden no estar disponibles por separado y el sistema puede incluir un módulo de diarización de flujo para separar al menos dos flujos de voz, pertenecientes a diferentes usuarios, de la comunicación por voz proporcionada.

El sistema también puede comprender los siguientes módulos en el generador de marcadores:

- un analizador de señal de voz para medir niveles de ruido en un flujo de voz y variaciones en el volumen, y compararlos con patrones de ruido y volumen configurados previamente;

- un analizador de prosodia para detectar la entonación del habla y compararla con patrones configurados previamente;

- un motor de reconocimiento de voz automático, configurado para la localización de palabras, para detectar en un flujo de voz cualquiera de las palabras de una lista definida previamente;

- un motor de reconocimiento de voz automático configurado para la transcripción, para medir la velocidad del habla en palabras por unidad de tiempo;

- un analizador de segundo orden conectado a los módulos previos para detectar marcadores repetidos, marcadores que se producen simultáneamente en ambos dichos al menos dos flujos de voz y demuestran cierta correlación comparándolos con patrones de correlación.

Y el procesador de acciones también puede comprender los siguientes módulos:

- un divisor de flujo de audio para extraer un segmento de un flujo de voz definido por su hora de inicio y su hora de finalización, o duración;

- un reconocimiento de voz de audio configurado para la transcripción de un flujo de voz en texto;

- un módulo de procesamiento de texto para buscar palabras, frases o patrones específicos.

Un último aspecto de la invención se refiere a un producto de programa de ordenador que comprende un código de programa de ordenador adaptado para realizar el procedimiento de la invención, cuando dicho código de programa se ejecuta en un ordenador, un procesador de señales digitales, una formación de compuertas programables en el terreno, un circuito integrado específico de la aplicación, un microprocesador, un micro-controlador o cualquier otra forma de hardware programable.

Descripción de los dibujos

Para completar la descripción que está realizándose, y con el objetivo de ayudar a entender mejor las características de la invención, según un ejemplo preferido de una realización práctica de la misma, acompañando a dicha descripción como parte integrante de la misma, hay un conjunto de dibujos en los que, a modo de ilustración y sin

limitación, se ha representado lo siguiente:

la **figura 1** muestra una muestra de una comunicación por voz entre 2 partes.

La **figura 2** muestra un diagrama de bloques que representa una realización de la invención.

La **figura 3** muestra otra realización que añade un módulo de diarización a la realización de la figura 2.

La **figura 4** muestra una realización del módulo generador de marcadores en detalle.

La **figura 5** muestra una realización del módulo procesador de acciones en detalle.

La figura 6 ilustra la generación de marcadores para una realización particular de la invención.

La figura 7 ilustra el procesador de acciones para la misma realización particular que la figura 6.

Descripción detallada de la invención

La invención describe un proceso, en términos generales, para analizar flujos de audio interrelacionados, tales como aquellos que proceden de llamadas de voz, para extraer, y convertir potencialmente en texto, la información relevante contenida en la comunicación por voz. A continuación se da a conocer en detalle un caso específico de comunicaciones por voz entre dos partes según una realización de la invención. Sin embargo, el procedimiento puede generalizarse para su aplicación a una comunicación genérica por voz entre n partes, de maneras evidentes para un experto en la técnica.

El procedimiento propuesto en esta invención identificaría información relevante en la conversación de voz y la presentaría al usuario. Esta invención no tiene como objetivo una transcripción textual completa, sino que se centra en identificar la información relevante en la conversación de voz.

Como etapa intermedia en el procedimiento, el procesamiento identifica *marcadores* en cada uno de los flujos de voz individuales. Los marcadores se definen como fragmentos con sello temporal en un flujo de voz de audio, que se marcan para ayudar en el proceso de reconocimiento. Se usan algunas características extraídas de los flujos de voz de audio para generar los marcadores, e incluso se incluyen. Es posible definir un marcador como una tupla que contiene las siguientes características, o campos, según esta realización:

- Hora de inicio
- Hora de finalización
- Identificador de canal (es decir, para una llamada de voz de dos partes, si está en el canal de la parte del que llama o el canal de la parte a la que se llama, o ambos)
- Tipo de marcador
- Detalles del marcador

El procesamiento individual de los flujos de voz puede dar como resultado la identificación de marcadores según diversos criterios, dando como resultado por tanto diferentes tipos de marcador. Según esta realización particular de la invención propuesta, los marcadores generados se dan a conocer más adelante. Una vez que se han elegido las características que van a buscarse, se comparan los flujos de voz con patrones configurados previamente, para identificar los marcadores:

- Marcadores de localización de palabras: estos marcadores se identifican cuando se detecta una palabra o una frase en uno de los flujos de voz de audio. Un ejemplo es un marcador para la detección de la frase “Deberías tomar nota de esto”. En este caso, el campo de “detalles del marcador” de la tupla puede contener la palabra o frase que se detectó.

- Marcadores de entonación: estos marcadores se identifican cuando se detecta un determinado patrón de entonación en uno de los flujos de audio. Un ejemplo puede ser un aumento considerable en la entonación. En este caso, el campo “detalles del marcador” de la tupla puede contener información sobre el patrón de entonación que se detectó.

- Marcadores de volumen: estos marcadores se identifican cuando el volumen del audio en un canal cambia de una manera específica. Un ejemplo puede ser un periodo sostenido de volumen de voz aumentado. En este caso, el campo “detalles del marcador” de la tupla puede contener información sobre el cambio en el volumen que se detectó.

- Marcadores de velocidad del habla: estos marcadores se identifican cuando la velocidad del habla en un canal cambia de una manera específica. Un ejemplo puede ser la detección de un intervalo con una velocidad del habla más lenta. En este caso, el campo “detalles del marcador” de la tupla puede contener información sobre la

manera específica en la que cambió la velocidad del habla.

- Marcadores de ruido: estos marcadores se identifican cuando se detectan determinados patrones de ruido o niveles de ruido en un canal. Un ejemplo puede ser un intervalo con música alta. En este caso, el tipo, nivel y patrón del ruido pueden estar incluidos en el campo “detalles del marcador” de la tupla.

- Marcadores de confianza de transcripción: estos marcadores se identifican cuando una palabra, o secuencia de palabras, tienen confianzas de transcripción por encima o por debajo de un determinado umbral, o dentro de un intervalo de confianza específico. Un ejemplo de esto puede ser la detección de una secuencia de al menos cinco palabras consecutivas con un nivel de confianza por debajo del 50%. La información de la palabra o las palabras, y sus confianzas respectivas, pueden incluirse en el campo “detalles del marcador” de la tupla.

- Marcadores mixtos: estos marcadores se identifican cuando se produce una combinación de otros marcadores simultáneamente, o de otro modo, en combinación. Un ejemplo puede ser la detección simultánea de la velocidad del habla más lenta y la confianza de transcripción baja. La información sobre los sucesos individuales que activaron el marcador puede incluirse en el campo “detalles del marcador” de la tupla.

Del mismo modo se realiza un procesamiento doble de los flujos de voz (es decir, un procesamiento simultáneo de ambos flujos de voz) y, de este modo, puede identificarse un segundo conjunto de marcadores según diversos criterios. Los marcadores generados por el procesamiento doble reflejan diferentes tipos de correlación entre los flujos de voz, dando como resultado por tanto diferentes tipos de marcador:

- **Marcadores de correlación**: estos marcadores se identifican cuando se cumple una determinada condición en ambos canales de audio, cuando cada flujo de voz de esta realización particular pertenece a un canal diferente. Un ejemplo de un marcador de correlación en relación con la localización de palabras clave puede ser “la frase ‹toma nota de esto› y la palabra ‹repetir› se dicen en diferentes canales en un intervalo de 5 segundos”. En este caso, los campos “detalles del marcador” de la tupla pueden incluir detalles sobre la condición de correlación que se cumplió.

- **Marcadores de repetición entre canales**: un subtipo específico de un marcador de correlación, los marcadores de repetición entre canales se identifican cuando se detecta la misma palabra o frase en ambos canales dentro de un determinado intervalo de tiempo. No se requiere una coincidencia exacta para identificar el marcador de repetición entre canales (es decir, puede haber una tolerancia a ligeras discrepancias entre las palabras dichas en un canal y en el otro). Un ejemplo es el siguiente diálogo:

Canal A: “Mi número es siete cuatro seis”

Canal B: “Siete cuatro seis…”

Canal A: “Dieciséis”

Canal B: “Uno seis…”

No se pretende que estas listas de tipos de marcador sean exhaustivas, sino que deben considerarse como una realización particular de la invención propuesta para complementar el sumario de la invención sin tener en cuenta ninguna limitación derivada de esta realización particular. Evidentemente pueden usarse diferentes criterios y sucesos de activación para la identificación de marcadores, sin desviarse del espíritu de la invención descrita.

Después de haber generado los marcadores, el procedimiento de la invención realiza determinadas acciones sobre los flujos de voz, según los marcadores, para extraer piezas de información potencialmente relevante de la comunicación por voz. En esta realización particular, las acciones que pueden adoptarse son:

- Extraer una pieza de información especificada en el marcador contenido dentro de los sellos temporales de inicio y finalización del marcador, y hacerla pasar a través de un motor de transcripción.

- Extraer la pieza de información, no del flujo de voz (o canal) especificado en el marcador, sino del otro canal, después del sello temporal de finalización del marcador, para una duración específica, y hacerla pasar a través de un motor de transcripción.

Como una variación, en otra realización, se extrae la pieza de información empezando por la primera vez que se detecta el habla, después del sello temporal del marcador.

- Extraer la pieza de información del otro canal (*no* el especificado en el marcador) después del sello temporal de finalización del marcador, para una duración específica, y hacerla pasar a través de un motor de ASR basado en la gramática, con una gramática específica.

- Extraer la pieza de información del otro canal (*no* el especificado en el marcador) después del sello temporal de finalización del marcador, para una duración específica, y hacerla pasar a través de un motor de ASR con un modelo de lenguaje (LM) específico.

Las acciones anteriores dan como resultado un fragmento de texto y/o audio extraído de la conversación de voz, que contiene una pieza de información relevante. Después de haber procesado todos los marcadores, la recopilación de esos fragmentos constituye una representación útil de la conversación de voz.

La figura 1 representa una muestra de una conversación de voz de 2 partes. El canal 1 (1) ilustra gráficamente un marcador identificado - específicamente, un marcador de localización de palabras de un único canal (3) para las palabras “puede repetir” - y una acción (4) asociada con el mismo. La acción en este ejemplo es: “ejecutar un fragmento de 6 segundos a través de un motor de transcripción”. El área de la señal de audio englobada dentro del área de marcador (5) representa el fragmento de audio comprendido entre los sellos temporales de inicio y finalización del marcador. La acción ejemplar ilustrada es la extracción de la pieza de información del otro canal (no el especificado en el marcador), del sello temporal de finalización del marcador y para una duración específica de seis segundos, seguida por su transcripción. La ventana de seis segundos se representa en la figura 1 mediante el área 5 dentro de la señal del canal 2 (2).

La figura 2 representa el sistema de la invención. Los flujos de audio de voz (21) se proporcionan a un módulo generador de marcadores (22), que procesa los flujos de audio de voz de la llamada de voz y genera los marcadores correspondientes. La entrada del generador de marcadores, como ya se mencionó, puede comprender varios flujos de audio de voz. Una vez generados los marcadores, dichos marcadores (23) se envían a un procesador de acciones (24) que toma los marcadores como entrada y ejecuta determinadas acciones sobre los flujos de audio de voz según cierta configuración, generando información relevante (25) como resultado. Las siguientes figuras profundizan más en detalle.

Como elemento adicional para la realización del sistema representado en la figura 2, la figura 3 representa un módulo de diarización del orador (37). La comunicación por voz (36) se dirige al módulo de diarización para solucionar los casos en los que los diferentes canales en la comunicación por voz no están disponibles por separado. Entonces el módulo de diarización del orador recibe la comunicación por voz en un fragmento y, aplicando un análisis de voz en la comunicación por voz, dicha comunicación por voz se divide en varios flujos, incluyendo cada uno las partes de la comunicación por voz de entrada en las que habla cada parte. A continuación, los flujos de audio de voz resultantes pueden alimentar al módulo generador de marcadores, que es equivalente al del sistema ilustrado en la figura 2.

La figura 4 representa una realización más detallada del módulo generador de marcadores (22). Aunque sólo se ilustra un flujo de audio de voz, debe entenderse que puede usarse el mismo ejemplo o un ejemplo diferente de cada uno de los componentes ilustrados para procesar cada uno de los flujos de audio. Por ejemplo, en el caso básico de una conversación de voz bilateral, en un caso el módulo de duplicación del flujo de audio puede actuar en el primer flujo, y en otro caso ese módulo puede actuar en el segundo flujo, etc.

Un módulo de duplicación del flujo de audio (41) duplica cada trama del flujo de voz entrante para generar N copias del flujo de voz, para su envío a los módulos procesadores de audio. Los módulos procesadores de audio son los módulos que procesan cada flujo de voz para generar los marcadores. En la figura 4, la realización representada comprende cuatro módulos procesadores de audio:

- Un analizador de señales de voz (42), que puede medir niveles de ruido en la señal de voz y variaciones en el volumen de voz, y compararlos con patrones de ruido (46) o patrones de volumen (47) configurados previamente. Este módulo genera marcadores de volumen y marcadores de ruido, tal como se describió anteriormente.

- Un analizador de prosodia (43), que puede detectar la entonación del habla, y compararla con patrones de entonación configurados previamente (48). El analizador de prosodia y el analizador de señales de voz son, en realizaciones alternativas de la invención, módulos separados, o sólo instancias de un detector de actividad de voz (VAD) disponible en el estado de la técnica.

- Un motor de ASR configurado para la localización de palabras (44), que toma una lista de palabras configuradas (49) y puede detectarlas en el flujo de audio de voz, si están presentes.

- Un motor de ASR configurado para la transcripción (45), que produce una transcripción del flujo de audio de voz y puede medir la velocidad del habla comparándola con patrones de velocidad del habla configurados previamente (50) (es decir, palabras por unidad de tiempo); calcular una confianza de la transcripción, comparándola con patrones de confianza de transcripción, configurados previamente (51), de cada palabra o frase; buscar construcciones específicas en el habla, más allá de frases fijas (tales como expresiones de fecha o números de teléfono). El motor de ASR configurado para la localización de palabras y el motor de ASR configurado para la transcripción son, en realizaciones alternativas de la invención, módulos separados, o sólo instancias de un motor de reconocimiento de voz (ASR), estando configurada cada instancia con diferentes modelos y/o modalidades de operación.

- Analizador de segundo orden (53). Siempre que se genere un marcador por cualquiera de los módulos de procesador de audio (42, 43, 44, 45), un analizador de segundo orden evalúa el conjunto de todos los marcadores generados (54). Este análisis puede dar como resultado la generación de uno o más marcadores de segundo orden, tales como marcadores de correlación si la comparación de los marcadores generados con patrones de correlación

configurados previamente (55) es satisfactoria, o marcadores de patrón mixtos si la comparación de los marcadores generados con patrones mixtos configurados previamente (56) es satisfactoria. Estos nuevos marcadores pueden activar de nuevo el procesado del analizador de segundo orden y, a su vez, dar como resultado la generación de marcadores adicionales. Aunque la configuración de los patrones mixtos y los patrones de correlación evita que este proceso caiga en bucles infinitos o tiempos de procesamiento prolongados, pueden fijarse restricciones adicionales para garantizar esto. Ejemplos de tales restricciones son: no tener en cuenta marcadores con una antigüedad de más de N segundos; tener en cuenta únicamente los M marcadores generados en último lugar.

Además, según una realización de la invención, el análisis de segundo orden también puede dar como resultado la eliminación de uno o más marcadores, según reglas configuradas. Un ejemplo de esto es evitar marcadores duplicados (por ejemplo dos marcadores de “volumen de voz alto” que aparezcan demasiado cerca en el tiempo entre sí), o cuando un nuevo marcador generado en el análisis de segundo orden hace que los marcadores de primer orden que lo activaron sean innecesarios (por ejemplo, un determinado marcador de patrón de entonación y un marcador de localización de palabras para el nombre del otro participante en la conversación [“¡Juan!”] pueden activar un marcador de segundo orden de “llamada de atención”, que transmite el significado compuesto de ambos marcadores de primer orden, haciendo que sean innecesarios).

La figura 5 ilustra una realización más detallada del módulo procesador de acciones (24). Los marcadores generados por el módulo generador de marcadores (22) son la entrada para el módulo procesador de acciones, que realiza acciones en los flujos de audio de voz según la información contenida en los marcadores y extrae la información relevante contenida en la comunicación por voz.

El módulo procesador de acciones está configurado para decidir qué acción corresponde a cada marcador. El módulo procesador de acciones toma cada marcador generado, calcula la acción a realizar según la configuración de acción (60), a continuación la ejecuta por medio de un proceso de ejecución de acciones. La realización particular de la figura 5 ilustra tres procesos diferentes de ejecución de acciones a modo de ejemplo:

- Un divisor de flujo de audio (57) para extraer un segmento de un flujo de audio definido por su sello temporal de inicio y su sello temporal de finalización (o duración).

- Un motor de ASR configurado para la transcripción (58), para transcribir la voz contenida en un flujo de audio de voz, o un segmento de audio, a texto.

- Un procesador de texto (59), para buscar palabras, frases o patrones (gramática) en un texto de entrada.

El procesamiento de un marcador puede dar como resultado una o más acciones; la ejecución de una acción puede implicar uno o más procesos de ejecución de acciones (por ejemplo, puede implicar la extracción de un segmento de 5 segundos de un flujo de audio, hacer pasar ese segmento a través del motor de ASR para su transcripción, a continuación hacer pasar la transcripción a través del procesamiento de texto para identificar un patrón de fecha). La ejecución de una acción puede dar como resultado cero o más fragmentos de información relevante.

Las figuras 6 y 7 presentan una realización particular de la invención propuesta. Para proporcionar algo de claridad, está implícita una definición simplificada de información relevante, concretamente, la información relevante sólo comprende lo que dicen las partes después de una frase introductoria tal como “por favor anote …”, “escriba …” o “puede tomar nota de …”. Además, esta realización particular considera una llamada normal, entre 2 partes, en la que ambos canales están disponibles por separado para el procesamiento.

Específicamente, la figura 6 ilustra la generación de marcadores para esta realización particular. Hay un módulo procesador de audio, concretamente, un motor de ASR para la localización de palabras (70). En la figura 6, están activas dos instancias del módulo y se encargan de procesar cada uno de los flujos de audio en la llamada de voz. El motor de ASR para la localización de palabras está configurado para etiquetar cada incidencia de las frases de una lista de localización de palabras (71): “anote”, “escriba”, “tome nota”. En la figura 6, se supone que la persona que llama produce el flujo de voz 74 y dice una de estas frases en dos momentos diferentes de la conversación de voz, por tanto, se generan los marcadores 1 y 3 (72). La persona a la que se llama produce el flujo de voz 75 y dice una de esas frases una vez durante la conversación de voz (por tanto, se genera el marcador 2 (73))

La figura 6 también muestra la definición específica del marcador 1 (76), como una tupla que contiene la hora de inicio y la hora de finalización, el canal al que se refiere (en este caso, el canal 1, que es la identificación dada a la persona que llama), el tipo de marcador (localización de palabras) y los detalles específicos del marcador. Para un marcador de localización de palabras, se supone que se incluyen tanto la frase específica que activó el marcador (en este caso, “tome nota”) como la confianza de la detección.

La figura 7 ilustra el módulo procesador de acciones para esta realización particular. Como continuación de la figura 6, que ilustra el generador de marcadores de la realización, la salida (76) es la entrada del procesador de acciones. El procesador de acciones comprende dos procesos de ejecución de acciones: un divisor de flujo de audio (77) y un motor de ASR para la transcripción (78). También se muestra una configuración de acción de muestra, que contiene una regla (79). Esta regla especifica las acciones que van a realizarse en un marcador de localización de palabras como, por ejemplo, el “marcador 1” (76), concretamente:

- Extraer un segmento del canal al que se refiere el marcador, comenzando 3 segundos después de la hora de finalización del marcador, y con una duración de 15 segundos.

- Transcribir ese segmento.

Esta regla pretende recuperar la información específica que el orador pretendía destacar cuando dijo la frase de activación, que habitualmente viene un poco después de decir “anote que”, “escriba”, etc.

Después de haber aplicado la acción, se obtiene un texto resultante, que contiene la información relevante identificada (80).

Un ejemplo de un caso para ello puede ser:

Para la regla:

“Regla de configuración de acción nº 1:

Si tipo _ marcador = “localización de palabras”:

Segmento=divid(id_flujo,hora_fin +3s, 15s)

Transcribir (segmento)”

Y para la información relevante:

“detalles mi número de teléfono es uno ocho tres cuatro nueve dos seises nueve tres cuatro cinco. Lo repetiré”

Como puede observarse, el sistema proporciona información útil, aunque la calidad de la información identificada puede mejorarse en este caso añadiendo, según otra realización de la invención, un módulo de procesamiento de texto como un proceso de ejecución de acciones, configurado para buscar números de teléfono en un texto de entrada.

El procedimiento de la invención, según una realización particular, comprende procesar cada uno de los flujos de voz en la comunicación por voz de manera consecutiva o, según otra realización, en paralelo. Puede realizarse en tiempo real (es decir, mientras se produce la comunicación por voz), o tras su finalización.

A lo largo de todo este documento, se usan los términos llamada de voz, conversación de voz y comunicación por voz de manera intercambiable para significar la comunicación síncrona entre dos o más partes por voz. Todo lo que se indica sobre llamadas de voz puede aplicarse a comunicaciones de vídeo, considerando sólo la parte de voz de la comunicación de vídeo.

REIVINDICACIONES

1. Procedimiento para obtener información relevante de una comunicación por voz proporcionada entre al menos dos usuarios, en el que la comunicación por voz comprende al menos dos flujos de voz (21) pertenecientes a dichos al menos dos usuarios, comprendiendo el procedimiento las siguientes etapas:

a) generar marcadores con sellos temporales (23), asociados a dichos al menos dos flujos de voz, según determinadas características extraídas de dichos al menos dos flujos de voz, en el que, cuando cierta característica es común (55) a ambos flujos de voz, generar un marcador con sello temporal (54) adicional, que indica una correlación entre dichos al menos dos flujos de voz (21);

b) extraer (57) determinadas piezas de información de dichos al menos dos flujos de voz, según cierta información que comprenden los marcadores;

c) transcribir (58) las determinadas piezas de información en texto para obtener la información relevante (25) de la comunicación por voz.

2. El procedimiento según la reivindicación 1, que comprende además separar dichos al menos dos flujos de voz, pertenecientes a dichos al menos dos usuarios, de la comunicación por voz.

3. El procedimiento según una cualquiera de las reivindicaciones anteriores, en el que las determinadas características de la etapa a) comprenden al menos una de las siguientes características: entonación, volumen, velocidad del habla, ruido.

4. El procedimiento según una cualquiera de las reivindicaciones anteriores, en el que la cierta información que comprenden los marcadores está contenida en una tupla que, a su vez, comprende elementos de la siguiente lista: un identificador para el flujo de voz perteneciente a cada usuario, un sello temporal de inicio, un sello temporal de finalización, una indicación del tipo de marcador, otros detalles del marcador.

5. El procedimiento según una cualquiera de las reivindicaciones anteriores, que comprende además comparar una característica de dichos al menos dos flujos de voz con un patrón configurado previamente.

6. El procedimiento según la reivindicación 4, en el que la etapa b) de extraer una pieza de información comprende además extraer la pieza de información contenida dentro de los sellos temporales de inicio y finalización definidos en el marcador.

7. El procedimiento según la reivindicación 4, en el que la etapa b) de extraer una pieza de información comprende además extraer la pieza de información, del otro flujo de voz distinto al indicado por el identificador de la tupla, contenida después del sello temporal de inicio definido en el marcador para una duración específica.

8. El procedimiento según una cualquiera de las reivindicaciones anteriores, que comprende además pasar las piezas de información extraídas a través de un sistema de reconocimiento de voz automático, basado en la gramática, con una gramática específica.

9. El procedimiento según una cualquiera de las reivindicaciones anteriores, que comprende además pasar las piezas de información extraídas a través de un sistema de reconocimiento de voz automático con un modelo de lenguaje específico.

10. El procedimiento según cualquiera de las reivindicaciones anteriores, en el que las etapas se realizan en tiempo real.

11. El procedimiento según cualquiera de las reivindicaciones 1 a 9, que comprende además almacenar la comunicación por voz en un dispositivo de almacenamiento de datos.

12. Sistema para obtener información relevante de una comunicación por voz proporcionada entre al menos dos usuarios, en el que la comunicación por voz comprende al menos dos flujos de voz, pertenecientes a dichos al menos dos usuarios, comprendiendo el sistema:

- un generador de marcadores (22) que está adaptado para recibir dichos al menos dos flujos de voz (21) que van a analizarse, y está adaptado para generar marcadores con sellos temporales, según ciertas características extraídas desde dichos al menos dos flujos de voz, en el que, cuando cierta característica es común (55) a ambos flujos de voz, está adaptado para generar un marcador con sello temporal (54) adicional, que indica una correlación entre dichos al menos dos flujos de voz;

- un procesador de acciones (24) adaptado para extraer determinadas piezas de información de dichos al menos dos flujos de voz, según alguna información comprendida por los marcadores, y adaptado para transcribir las determinadas piezas de información, en texto, estando por tanto adaptado para obtener la información relevante (25) de la comunicación por voz.

13. Sistema según la reivindicación 12, que comprende además un módulo de diarización de flujo, adaptado para separar al menos dos flujos de voz, pertenecientes a diferentes usuarios, de la comunicación por voz proporcionada.

14. Sistema según la reivindicación anterior, en el que el generador de marcadores comprende además al menos uno de los siguientes módulos:

- un analizador de señal de voz, adaptado para medir niveles de ruido en un flujo de voz y variaciones en el volumen, y adaptado para compararlos con patrones de ruido y volumen configurados previamente;

- un analizador de prosodia adaptado para detectar la entonación del habla y compararla con patrones configurados previamente;

- un motor de reconocimiento de voz automático, configurado para la localización de palabras, adaptado para detectar en un flujo de voz cualquiera de las palabras o frases de una lista definida previamente;

- un motor de reconocimiento de voz automático, configurado para la transcripción, para medir una velocidad del habla en palabras por unidad de tiempo;

- un analizador de segundo orden conectado a los módulos previos, adaptado para detectar marcadores repetidos, marcadores que se producen simultáneamente en ambos de dichos al menos dos flujos de voz y demuestran cierta correlación comparándolos con patrones de correlación.

15. Sistema según la reivindicación 12, en el que el procesador de acciones comprende además los siguientes módulos:

- un divisor de flujo de audio, adaptado para extraer un segmento de un flujo de voz definido por su hora de inicio y su hora de finalización, o su duración;

- un módulo de reconocimiento de voz de audio, configurado para la transcripción de un flujo de voz en texto;

- un módulo procesador de texto, adaptado para buscar palabras, frases o patrones específicos.

16. Un producto de programa de ordenador que comprende un código de programa de ordenador adaptado para realizar el procedimiento según cualquiera de las reivindicaciones 1 a 11 cuando dicho código de programa se ejecuta en un ordenador, un procesador de señales digitales, una formación de compuertas programables en el terreno, un circuito integrado específico de la aplicación, un microprocesador, un micro-controlador o cualquier otra forma de hardware programable.

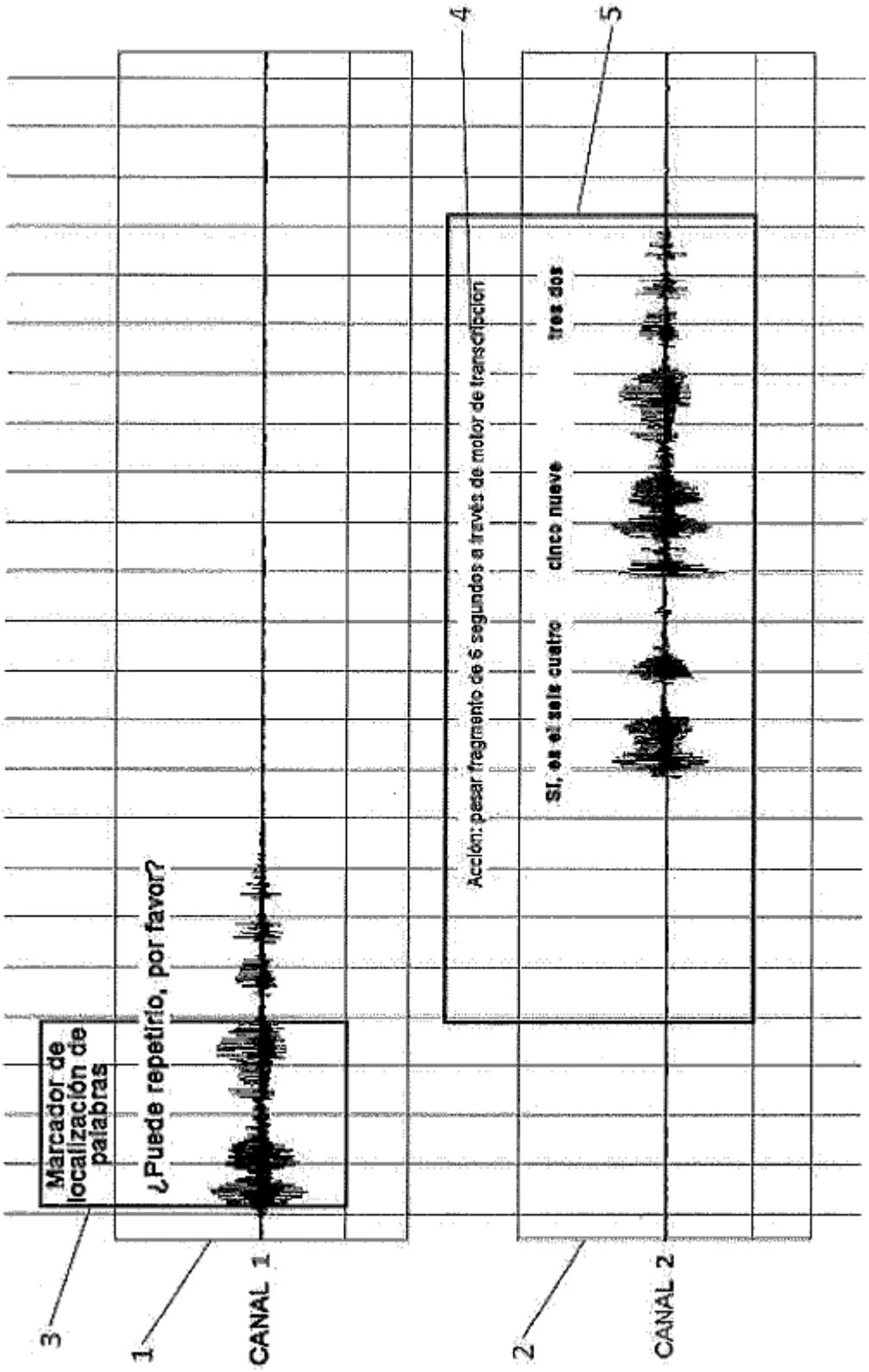
Marcador de localización de palabras
¿Puede repetirlo, por favor?
Acción: pesar fragmento de 6 segundos a través de motor de transcripción
Sí, es el seis cuatro cinco nueve tres dos
3
1
CANAL 1
4
5
2
CANAL 2

FIG. 1

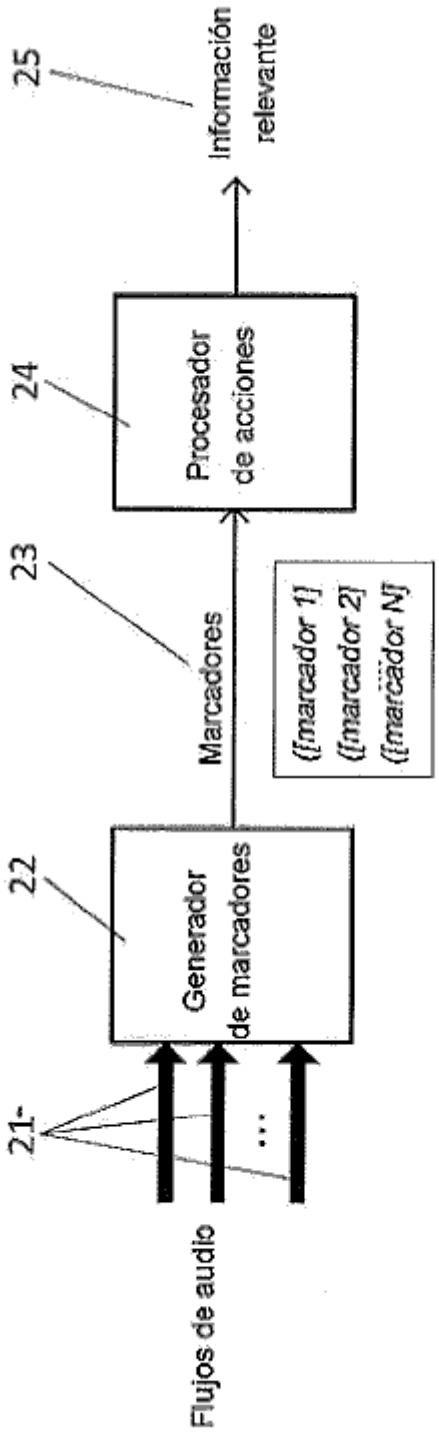
25
Información relevante
24
Procesador de acciones
23
Marcadores
{[marcador 1]
{[marcador 2]
{[marcador N]
22
Generador de marcadores
21-
...
Flujos de audio

FIG. 2

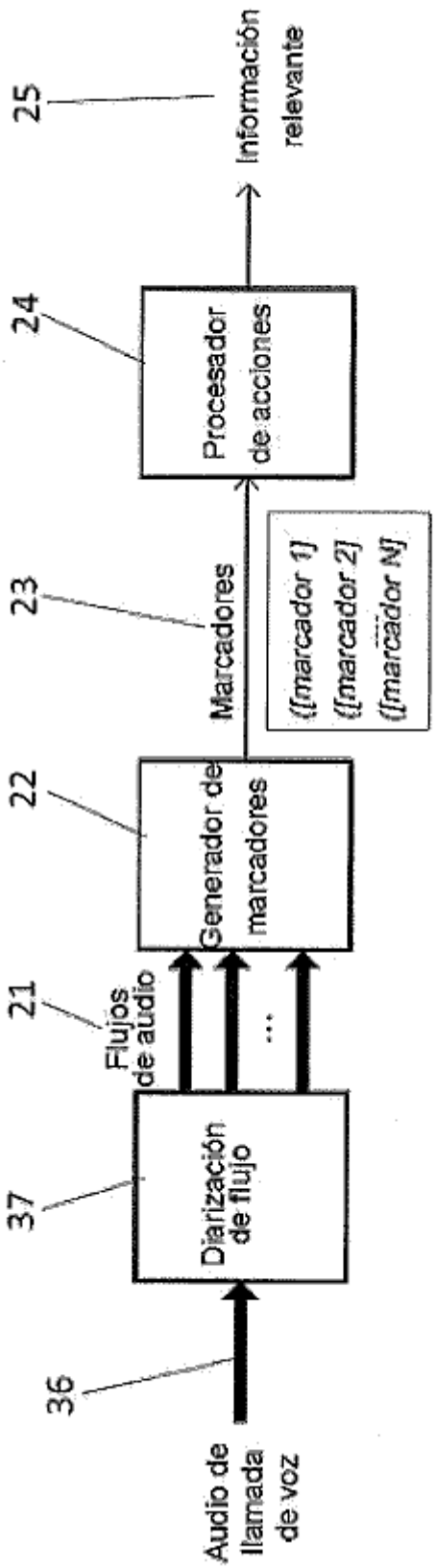
Audio de llamada de voz
36
37
Diarización de flujo
21
Flujos de audio
...
22
Generador de marcadores
23
Marcadores
{[marcador 1]
{[marcador 2]
{[marcador N]
24
Procesador de acciones
25
Información relevante

FIG. 3

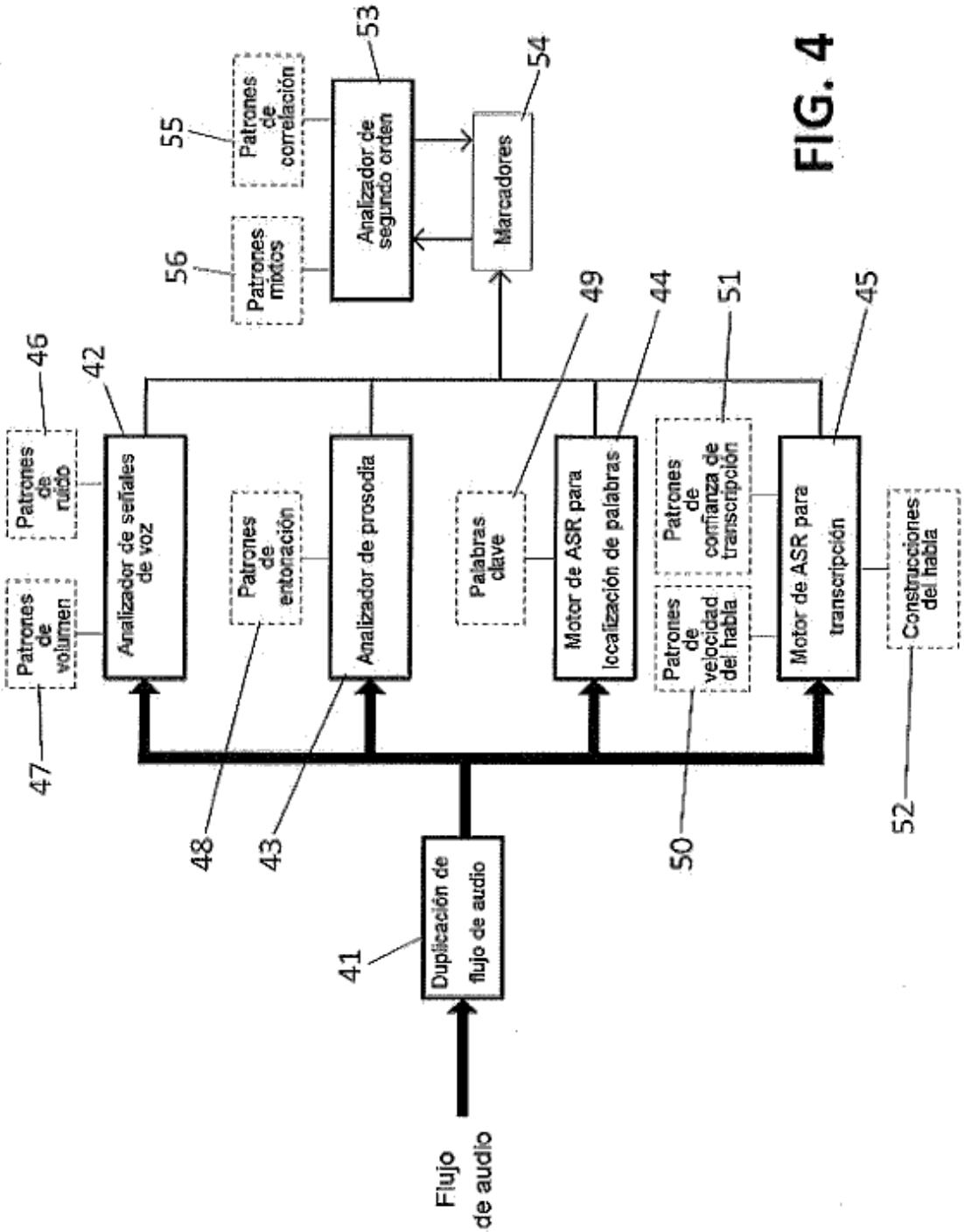
Flujo de audio
Duplicación de flujo de audio
41
Analizador de señales de voz
42
Analizador de prosodia
43
Motor de ASR para localización de palabras
44
Motor de ASR para transcripción
45
Patrones de ruido
46
Patrones de volumen
47
Patrones de entonación
48
Palabras clave
49
Patrones de velocidad del habla
50
Patrones de confianza de transcripción
51
Construcciones del habla
52
Analizador de segundo orden
53
Marcadores
54
Patrones de correlación
55
Patrones mixtos
56

FIG. 4

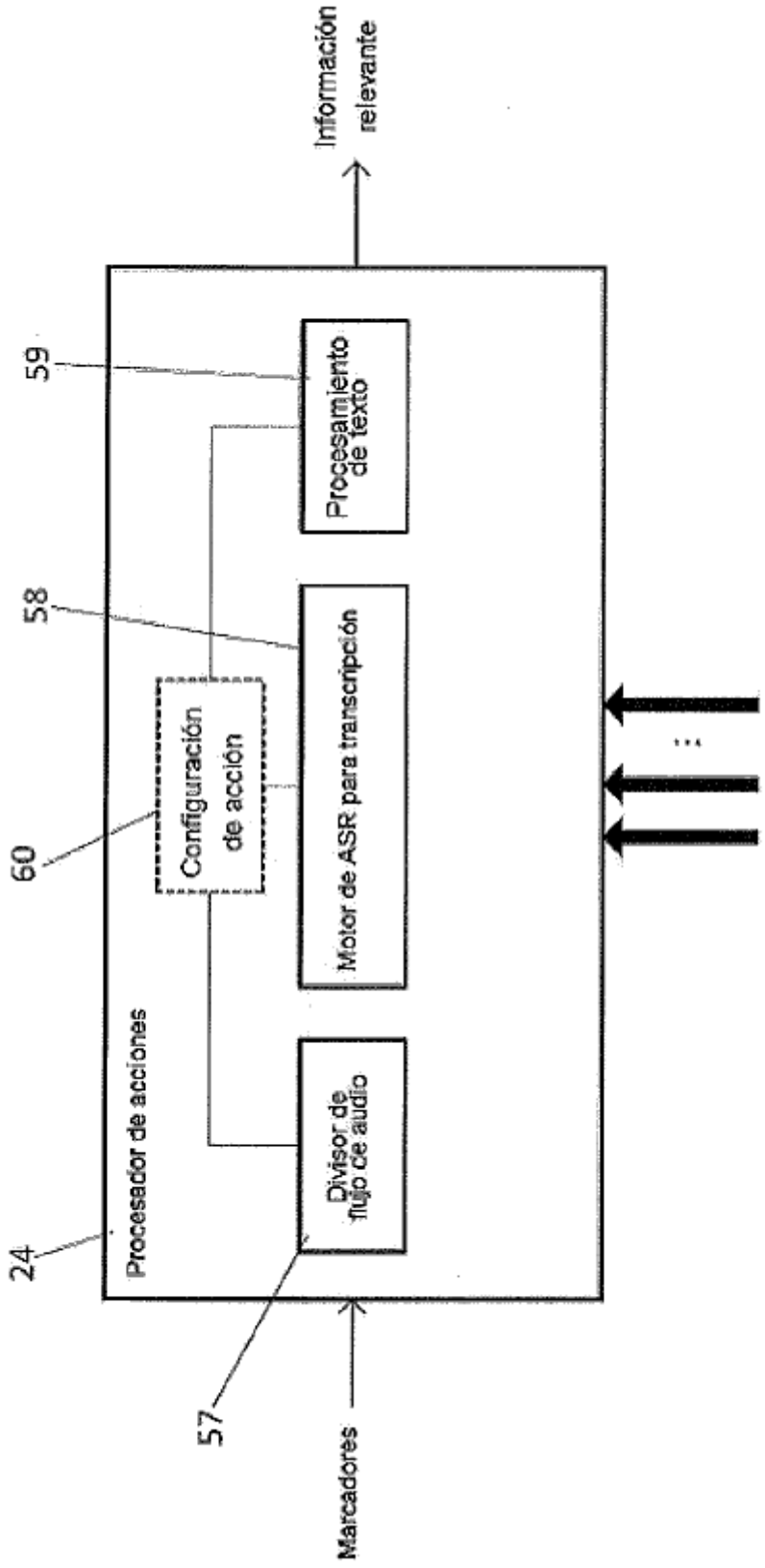
Información relevante
59
Procesamiento de texto
58
Motor de ASR para transcripción
60
Configuración de acción
Procesador de acciones
24
Divisor de flujo de audio
57
Marcadores
Flujos de audio

FIG. 5

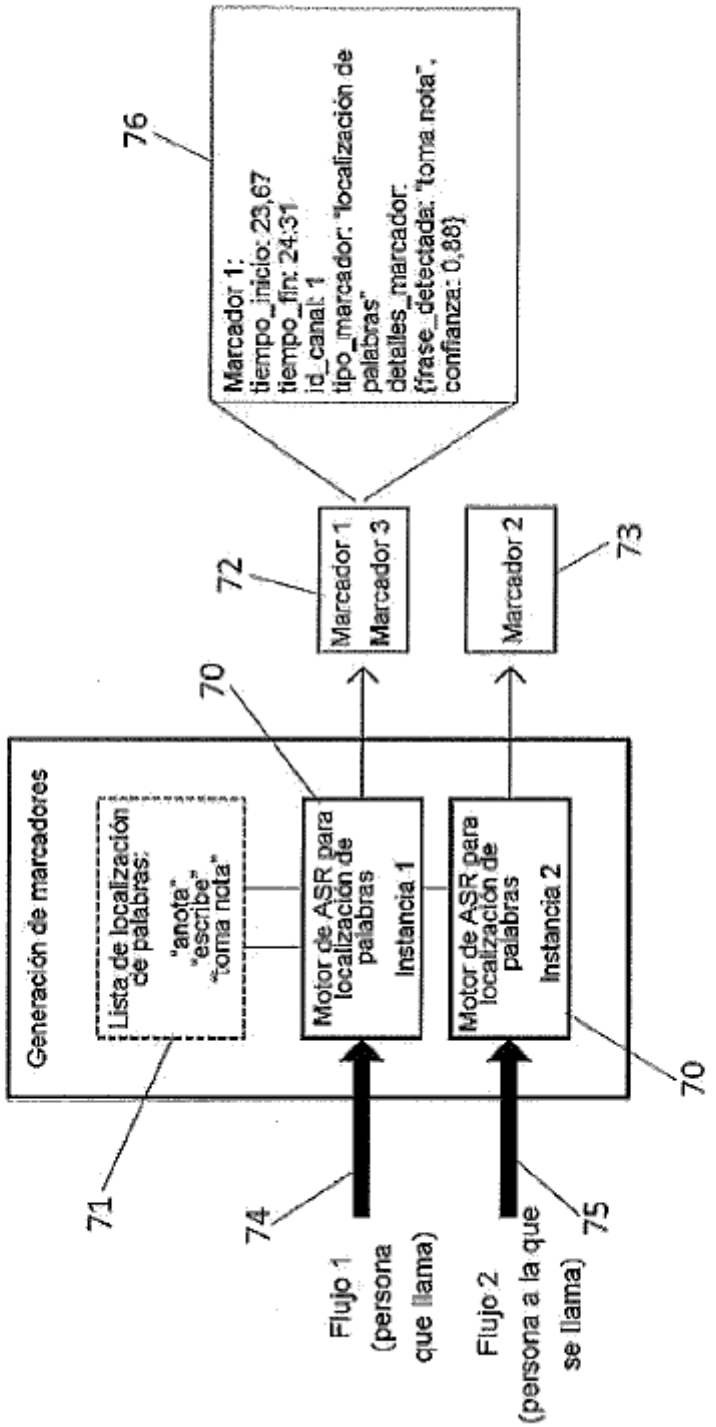
Generación de marcadores
Lista de localización de palabras:
"anota"
"escribe"
"toma nota"
Motor de ASR para localización de palabras
Instancia 1
Motor de ASR para localización de palabras
Instancia 2
Marcador 1
Marcador 3
Marcador 2
Marcador 1:
tiempo_inicio: 23,67
tiempo_fin: 24:31
id_canal: 1
tipo_marcador: "localización de palabras"
detalles_marcador:
{frase_detectada: "toma nota",
confianza: 0,88}
Flujo 1 (persona que llama)
Flujo 2 (persona a la que se llama)
70
71
72
73
74
75
76

FIG. 6

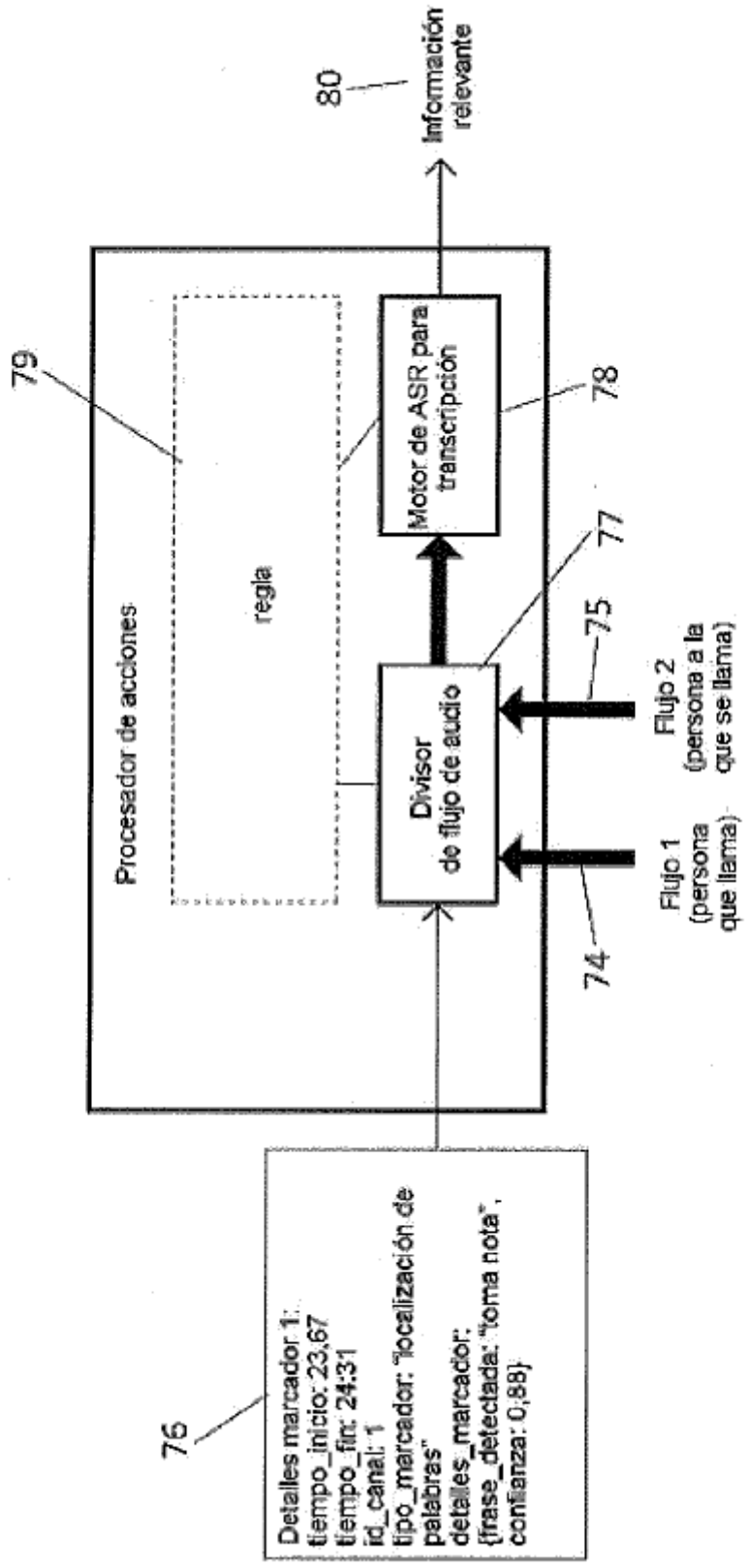
80
Información relevante
79
Procesador de acciones
regla
Motor de ASR para transcripción
78
77
75
Flujo 2 (persona a la que se llama)
Divisor de flujo de audio
Flujo 1 (persona que llama)
74
76
Detalles marcador 1:
tiempo_inicio: 23,67
tiempo_fin: 24:31
id_canal: 1
tipo_marcador: "localización de palabras"
detalles_marcador:
{frase_detectada: "toma nota",
confianza: 0,88}

FIG. 7