(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 318 589**

(51) Int. Cl.:
H04L 29/06 (2006.01)
G10L 15/28 (2006.01)
G10L 11/02 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **05850633 .8**
(96) Fecha de presentación : **28.12.2005**
(97) Número de publicación de la solicitud: **1847088**
(97) Fecha de publicación de la solicitud: **24.10.2007**

(54) Título: **Procedimiento de transmisión de marcas de fin de voz en un sistema de reconocimiento de voz.**

(30) Prioridad: **04.02.2005 FR 05 50322**

(45) Fecha de publicación de la mención BOPI:
01.05.2009

(45) Fecha de la publicación del folleto de la patente:
01.05.2009

(73) Titular/es: **France Télécom**
6, place d'Alleray
75015 Paris, FR

(72) Inventor/es: **Ferrieux, Alexandre**

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

DESCRIPCIÓN

Procedimiento de transmisión de marcas de fin de voz en un sistema de reconocimiento de voz.

La presente invención se refiere a un procedimiento de transmisión de marcas de fin de voz en un sistema de reconocimiento distribuido de voz que funciona en modo de transmisión discontinua.

La invención tiene una aplicación particularmente ventajosa en el ámbito general del reconocimiento de voz.

Más especialmente, el contexto de la invención es el del reconocimiento distribuido de voz o DSR por sus siglas en inglés "Distributed Speech Recognition", como se define en las normas ETSI ES 201 108, ES 202 050, ES 202 212 y el documento IETF RFC3557.

De una manera general, los procedimientos de reconocimiento de voz implican, en una primera fase, la extracción de parámetros acústicos extraídos de un segmento de voz pronunciado por un locutor, que puede ser el usuario de un terminal, entre otros de un teléfono móvil. En una segunda fase, un sistema especializado de reconocimiento de voz trata los parámetros acústicos obtenidos, con el objeto de restablecer el contenido fonético del segmento de voz pronunciado. Un servidor, que integra este sistema de reconocimiento de voz, puede entonces reaccionar a las palabras reproducidas de esta manera por el locutor. Este servidor es, por ejemplo, un servidor de voz en un sistema de telefonía móvil.

El reconocimiento distribuido de voz (DSR) consiste en efectuar las primeras etapas del reconocimiento de voz, es decir, la extracción de los parámetros acústicos, en el terminal mismo, y en transmitir sólo el resultado al servidor. Habiéndose elegido estos parámetros para optimizar los resultados del reconocimiento de voz, se obtiene, con velocidad equivalente a la de un codificador-decodificador (en inglés "codec") clásico para la conversación entre humanos, una mejora clara del reconocimiento.

El documento RFC3557 mencionado anteriormente describe la transmisión de los parámetros acústicos como carga útil del protocolo conocido por el acrónimo RTP (por sus siglas en inglés "Real Time Protocol", RFC3550). Un modo de aplicación de DSR propuesto en este documento se refiere a la Transmisión Discontinua, o TD, que consiste en que el terminal envía datos en dirección al servidor no permanentemente, sino sólo durante segmentos de voz. Con este fin, la transmisión de datos sólo se efectúa cuando el usuario pulsa una tecla de un dispositivo "Push-to-Talk" o bajo el control de un detector de actividad vocal (DAV). Se entiende que el interés de este modo de transmisión discontinua reside en el ahorro de ancho de banda durante los periodos de silencio.

Por supuesto, cuando se utiliza el modo TD, es necesario para el servidor de voz, por ejemplo, conocer el fin de los segmentos de voz con el fin de poder indicar al sistema de reconocimiento de voz que todos los datos de parámetros acústicos se reciben y que puede efectuar las operaciones de reconocimiento y finalizar su resultado. Para ello, el documento RCF3557 prevé paquetes de datos especiales que contienen tramos nulos y que sirven como marcas de fin de voz.

Un inconveniente del modo TD es que en caso de pérdida de paquetes de tramos nulos en la red durante la transmisión de los datos, el servidor, al no ser informado ya del fin de los segmentos de voz, no puede dar ninguna orden de ejecución al sistema de reconocimiento de voz. El resultado es que el servidor no puede reaccionar a las palabras del usuario, que debe sufrir entonces largos e inaceptables periodos de espera.

Para remediar este inconveniente, se propone un mecanismo de temporización que consiste en provocar una reacción del servidor si alguna información de fin de segmento de voz no se recibe al final de un lapso de tiempo dado. No obstante, este tipo de mecanismo ciego es necesariamente lento, debido a que está relacionado con los lapsos, a veces largos, de los segmentos de voz en una conversación normal.

Asimismo, el problema técnico que debe resolver el objeto de la presente invención es proponer un procedimiento de transmisión de marcas de fin de voz en un sistema de reconocimiento distribuido de voz que funciona en modo de transmisión discontinua, sistema en el que se emiten segmentos de voz, seguidos de periodos de silencio, terminando cada segmento de voz en una marca de fin de voz, que permitiría darle al canal de señalización constituido por las marcas de fin de voz una mejor solidez de cara a las pérdidas de transmisión que la obtenida con un mecanismo de temporización, garantizando lapsos relacionados con las únicas condiciones de red y que no se han fijado arbitrariamente a duraciones de temporización obligatoriamente más largas.

La solución al problema técnico planteado consiste, según la presente invención, en que dicha marca de fin de voz se reemita continuamente durante toda la duración del periodo de silencio que sigue a dicho segmento de voz.

De esta manera, incluso si se produce una pérdida de transmisión al final de un segmento de voz, provocando la pérdida de la marca de fin contenida en el segmento truncado, la información de fin de segmento podrá, no obstante, ser comunicada al servidor en el momento que la red vuelva a estar operativa ya que el servidor podrá recibir entonces la marca de fin reemitida después de retomarse la transmisión. Se notifica, por tanto, al servidor del fin del segmento con una reactividad muy fuerte, bien para ordenar la ejecución de la operación de reconocimiento, bien, por el contrario, para rechazar un segmento truncado por las pérdidas en línea.

El ritmo de reemisión de las marcas de fin de voz, es decir, la duración del intervalo de tiempo que separa dos marcas consecutivas reemitidas debe responder al siguiente compromiso:

- si es demasiado lento, el usuario puede sufrir fuertes latencias, añadiéndose de esta manera a los inconvenientes de los mecanismos de temporización anteriormente citados,

- si es demasiado rápida, el ancho de banda consumido durante los periodos de silencio puede acercarse a la de los periodos de voz, anulando entonces el interés de la transmisión discontinua TD. Además, esta rapidez puede ser inútil debido a la tolerancia temporal del usuario y la correlación temporal de las pérdidas de paquetes según la cual dos marcas de fin de voz reemitidas demasiado cerca tienen muchas posibilidades de perderse al mismo tiempo.

Son posibles dos opciones: según una primera opción, dicha marca de fin de voz se reemite en intervalos de tiempo de la misma duración, mientras que según una segunda opción, dicha marca de fin de voz se reemite en intervalos de tiempo de duración creciente. Esta segunda opción es ventajosa en términos de banda ancha, pero presenta el riesgo de reintroducir fuertes latencias.

Un compromiso satisfactorio consiste, según la invención, en que dicha duración es del orden del segundo.

En un modo de realización particular de la invención, se prevé que la reemisión de dicha marca de fin de voz se interrumpa al recibir un mensaje de acuse de recibo de una marca de fin de voz reemitida.

Esta disposición presenta la ventaja de un ahorro en banda ancha y se le dará por tanto preferencia cuando el ancho de banda disponible esté limitado. En el caso contrario, un acuse de recibo por parte del servidor no será necesario, considerándose el ancho de banda consumido tolerable aunque la primera marca de fin de voz llegue al servidor, aunque la reemisión de marcas de fin suplementarias se vuelve inútil.

Con el objeto de seguir limitando el consumo de banda ancha, la invención prevé que las marcas de fin de voz se transmitan en paquetes de longitud inferior a la longitud nominal de los pares de tramos en dichos segmentos de voz.

Por último, hay que señalar otra ventaja de la invención, particularmente apreciable durante fuertes pérdidas de transmisión. En efecto, si la red está muy perturbada, se puede producir una pérdida total de un segmento de voz. Si, además, la transmisión se restablece durante el periodo de silencio que sigue al segmento perdido, el servidor de voz podrá, no obstante, recibir una marca de fin de voz debido a la emisión continua de marcas de fin de voz prevista por la invención. Ahora bien, los paquetes que transportan estas marcas comprenden generalmente una indicación de la fecha de fin de voz del segmento considerado, de forma que comparando las fechas de las dos últimas marcas de fin de voz recibidas sucesivamente, el servidor puede detectar la pérdida del segmento de voz y reaccionar con respecto al usuario de manera apropiada, pidiéndole, por ejemplo, repetir el mensaje.

La presente invención se refiere también, según la reivindicación 7, a un terminal de un sistema de reconocimiento distribuido de voz que funciona en modo de transmisión discontinua, siendo apto el terminal para emitir segmentos de voz, seguidos de periodos de silencio, terminando cada segmento de voz en un marca de fin de voz, siendo además este terminal apto para reemitir dicha marca de fin de voz de forma continua durante toda la duración del periodo de silencio que sigue a dicho segmento de voz.

La invención se refiere también, según la reivindicación 13, a un sistema de reconocimiento distribuido de voz que funciona en modo de transmisión discontinua, que comprende al menos un terminal según la invención en el que la reemisión de la marca de fin de voz se interrumpe al recibir un mensaje de acuse de recibo de una marca de fin de voz reemitida y que comprende un servidor de voz apto para emitir un mensaje de acuse de recibo de una marca de fin de voz reemitida.

La descripción que aparece a continuación en relación con los dibujos anexos, dados a título de ejemplo no limitativo, hará comprender correctamente en qué consiste la invención y cómo se puede realizar.

La figura 1a es un esquema que muestra las operaciones efectuadas en un terminal aplicando el procedimiento según la invención.

La figura 1b es un esquema que muestra las operaciones efectuadas en un servidor de reconocimiento de voz asociado a un terminal de la figura 1a.

En la figura 1a están representadas las diferentes operaciones sucesivas efectuadas en un terminal, por ejemplo un teléfono móvil, en el marco general de un sistema de reconocimiento distribuido de voz donde un servidor de voz ilustrado en la figura 1b debe identificar los mensajes pronunciados por el usuario en el terminal.

Según la figura 1a, el mensaje de voz 10 emitido por el usuario se trata en el terminal mismo, según el procedimiento de reconocimiento distribuido de voz DSR. Este tratamiento se efectúa, por lo tanto, en una unidad 20 del terminal que comprende un módulo 21 que permite extraer de la señal sorda 10 los parámetros acústicos necesarios para el sistema de reconocimiento de voz del servidor para reconstituir el mensaje pronunciado por el usuario. Los métodos

de extracción de parámetros acústicos son muy conocidos y no forman parte del objeto de la presente invención. Sólo se recordarán las normas ETSI correspondientes: ES 201 108, ES 202 050 y ES 202 212.

Como indica la figura 1a, la operación de extracción de los parámetros acústicos por la aplicación de un modo de transmisión discontinua TD efectuada por un módulo 22 de la unidad 20 de tratamiento con el objetivo de limitar a los únicos segmentos de voz el envío de datos hacia el servidor. Con tal fin, el módulo 22 recibe de un indicador 23 una señal de comienzo de voz. Dicho indicador 23 puede ser un dispositivo "Push-to-Talk" donde el usuario pulsa una tecla cuando comienza a hablar o un detector de actividad vocal DAV.

La señal proporcionada por la unidad 20 de tratamiento del terminal está, de esta manera, constituida por segmentos 30, 40 de voz que comprenden paquetes que transportan en su carga útil parámetros acústicos extraídos por el módulo 21. Cada segmento de voz termina en una marca 31, 41 de fin de voz. Los dos segmentos 30 y 40 de voz consecutivos están separados por un periodo 34 de silencio.

Se puede ver en la figura 1a que la marca 31 de voz asociada al segmento 30 se reemite continuamente durante toda la duración del periodo 34 de silencio que sigue a dicho segmento. Las marcas de fin de voz reemitidas tienen como referencia 31a, 31b, etc.

El interés de una disposición como esta aparece claramente en la figura 1b en la que se representa un sistema 50 de reconocimiento de voz de un servidor de voz.

La señal que contiene los parámetros acústicos del usuario se transmite a través de la red hasta el sistema 50 que efectúa las operaciones de reconstitución del mensaje de voz pronunciado por el usuario a partir de los datos recibidos en los segmentos 30, 40 de voz. La marca 31 de fin de voz sirve para indicar al sistema 50 que el segmento 30 termina y que puede efectuar la operación de reconocimiento para ese segmento.

Si, como se indica en la figura 1b, la transmisión a través de la red se ha perturbado durante una duración T truncando de esta manera el fin del segmento 30 y, por ejemplo, las marcas 31 y 31a de fin de voz, el sistema 50 detectará la marca 31b, inmediatamente consecutiva a la reanudación de la transmisión. La operación de reconocimiento podrá efectuarse entonces de manera precoz, siendo el retraso introducido del orden de la duración de las pérdidas de red, por lo tanto, seguramente más corto que con los mecanismos de temporización utilizados habitualmente.

En las figuras 1a y 1b, dicha marca 31 de fin de voz se reemite en intervalos de tiempo de la misma duración Δt, por ejemplo, del orden del segundo. Pero, también se puede contemplar que la duración de los intervalos de tiempo que separan dos reemisiones consecutivas sea creciente, en una progresión de factor 1,5 ó 2 por ejemplo.

Como se ha indicado anteriormente, la emisión de las marcas 31, 31a,... de fin de voz puede interrumpirse al recibir el terminal un mensaje de acuse de recibo por el servidor de una marca de fin de voz. De esta manera, en el ejemplo de las figuras 1a y 1b, después de haber recibido la marca 31b, el servidor puede enviar al terminal un mensaje de acuse de recibo de esta marca. El terminal informado de esta manera podrá interrumpir el envío de nuevas marcas 31c, 31d,... de fin de voz que se vuelven inútiles.

Por último, se pueden realizar ahorros de ancho de banda limitando los paquetes que transportan las marcas 31a, 31b,... de fin de voz al mínimo necesario, de manera que su longitud sea notablemente inferior a la longitud nominal de los pares de tramos en los segmentos de voz.

REIVINDICACIONES

1. Procedimiento de transmisión de marcas de fin de voz en un sistema de reconocimiento distribuido de voz que funciona en modo de transmisión discontinua, sistema en el que se emiten segmentos (30, 40) de voz, seguidos de periodos (34) de silencio, terminando cada segmento (30, 40) en una marca (31) de fin de voz, **caracterizado** porque dicha marca (31) de fin de voz se reemite continuamente (31a, 31b, 31c, 31d) durante toda la duración del periodo (34) de silencio que sigue a dicho segmento (30) de voz.

2. Procedimiento según la reivindicación 1, **caracterizado** porque dicha marca (31) de fin de voz se reemite en intervalos de tiempo de la misma duración (Δt).

3. Procedimiento según la reivindicación 1, **caracterizado** porque dicha marca de fin de voz se reemite en intervalos de tiempo de duración (Δt) creciente.

4. Procedimiento según una de las reivindicaciones 2 ó 3, **caracterizado** porque dicha duración (Δt) es del orden del segundo.

5. Procedimiento según una cualquiera de las reivindicaciones 1 a 4, **caracterizado** porque la reemisión de dicha marca (31) de fin de voz se interrumpe al recibir un mensaje de acuse de recibo de una marca (31b) de fin de voz reemitida.

6. Procedimiento según una cualquiera de las reivindicaciones 1 a 5, **caracterizado** porque las marcas (31, 31a, 31b, 31c, 31d) de fin de voz se transmiten en paquetes de longitud inferior a la longitud nominal de los pares de tramos en dichos segmentos (30, 40) de voz.

7. Terminal de un sistema de reconocimiento distribuido de voz que funciona en modo de transmisión discontinua, siendo apto dicho terminal para emitir segmentos (30, 40) de voz, seguidos de periodos (34) de silencio, terminando cada segmento (30, 40) por una marca (31) de fin de voz, **caracterizado** porque dicho terminal es apto para reemitir dicha marca (31) de fin de voz continuamente (31a, 31b, 31c, 31d) durante toda la duración del periodo (34) de silencio que sigue a dicho segmento (30) de voz.

8. Terminal según la reivindicación 7, **caracterizado** porque dicha marca (31) de fin de voz se reemite en intervalos de tiempo de la misma duración (Δt).

9. Terminal según la reivindicación 7, **caracterizado** porque dicha marca de fin de voz se reemite en intervalos de tiempo de duración (Δt) creciente.

10. Terminal según una de las reivindicaciones 8 ó 9, **caracterizada** porque dicha duración (Δt) es del orden del segundo.

11. Terminal según una cualquiera de las reivindicaciones 7 a 10, **caracterizado** porque las marcas (31, 31a, 31b, 31c, 31d) de fin de voz se transmiten en paquetes de longitud inferior a la longitud nomina de los pares de tramos en dichos segmentos (30, 40) de voz.

12. Terminal según una cualquiera de las reivindicaciones 7 a 11, **caracterizado** porque la reemisión de dicha marca (31) de fin de voz se interrumpe al recibir un mensaje de acuse de recibo de una marca (31b) de fin de voz reemitida.

13. Sistema de reconocimiento distribuido de voz que funciona en modo de transmisión discontinua, que comprende al menos de un terminal según la reivindicación 12 y que comprende un servidor de voz apto para emitir un mensaje de acuse de recibo de una marca (31b) de fin de voz reemitida.

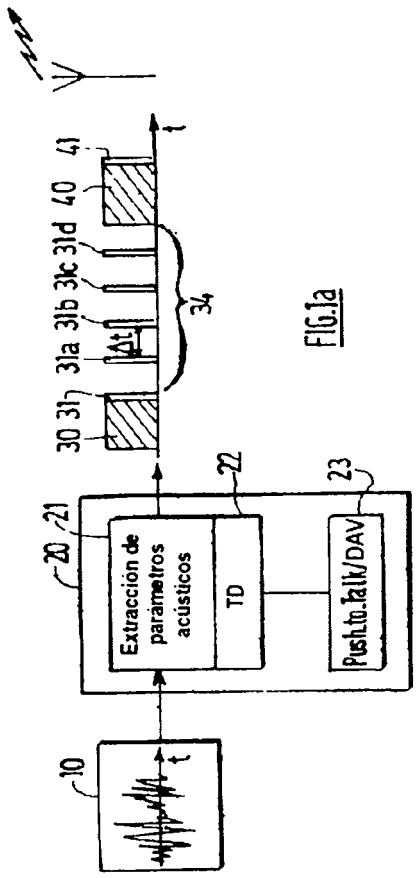
10
20
21
Extracción de parámetros acústicos
22
TD
23
Push to talk/DAV
30
31
31a
31b
31c
31d
34
40
41
t

FIG.1a

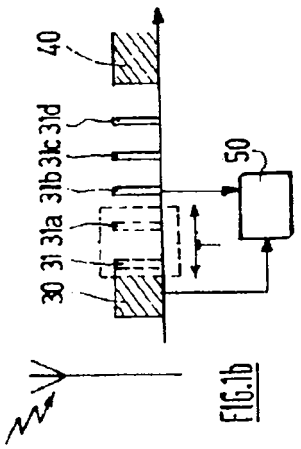
30
31
31a
31b
31c
31d
40
50

FIG.1b