(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 675 734**

(51) Int. CI.:

G10L 15/26 (2006.01)
G10L 25/78 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **06.04.2006 PCT/FR2006/050311**

(87) Fecha y número de publicación internacional: **12.10.2006 WO06106272**

(96) Fecha de presentación y número de la solicitud europea: **06.04.2006 E 06726317 (8)**

(97) Fecha y número de publicación de la concesión europea: **04.04.2018 EP 1866909**

(54) Título: **Procedimiento de sincronización entre una operación de procesamiento de reconocimiento vocal y una acción de activación de dicho procesamiento**

(30) Prioridad:

07.04.2005 FR 0550896

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
12.07.2018

(73) Titular/es:

ORANGE (100.0%)
78, rue Olivier de Serres
75015 Paris, FR

(72) Inventor/es:

MONNE, JEAN y
FERRIEUX, ALEXANDRE

(74) Agente/Representante:

ISERN JARA, Jorge

DESCRIPCIÓN

Procedimiento de sincronización entre una operación de procesamiento de reconocimiento vocal y una acción de activación de dicho procesamiento

La presente invención se refiere a un procedimiento de sincronización entre, por una parte, una operación de procesamiento por reconocimiento automático de la voz de una secuencia vocal emitida por un locutor y, por otra parte, al menos una acción de dicho locutor destinada a activar dicho procesamiento.

La invención encuentra una aplicación particularmente ventajosa en el campo del reconocimiento automático de la voz, principalmente en los terminales multimedia como terminales móviles de nueva generación, los asistentes personales (PDA) y los mandos a distancia que incluyen un micrófono.

Cuando está en comunicación con un servidor local equipado con un sistema de reconocimiento automático de la voz, un usuario de un terminal móvil, por ejemplo, no tiene en principio ninguna acción particular que efectuar para indicar al sistema de reconocimiento que va a enunciar una secuencia vocal. En efecto, el sistema está, o bien siempre a la escucha del usuario, o bien en condiciones de determinar, a partir de la estructura del diálogo entre el servidor y el usuario, los instantes en que este va a hablar.

En el caso en que permanece siempre a la escucha del usuario, el sistema de reconocimiento busca en el flujo sonoro continuo que recibe las zonas temporales susceptibles de corresponder a secuencias vocales emitidas por el usuario. Esta búsqueda se efectúa, de manera conocida, por medio de un dispositivo denominado de detección de actividad vocal. Por supuesto, para que este sistema funcione correctamente, es necesario que la detección de la actividad vocal no genere demasiadas falsas alertas o, en sentido contrario, que sean rechazadas por el mecanismo de reconocimiento automático de la voz.

Esto es por lo que la detección de la actividad vocal da los mejores resultados en un contexto de captura de sonido de proximidad, micrófono cerca de la boca del locutor, siendo favorecida la recepción de la voz del usuario en detrimento de la captura de los ruidos del entorno que perturban el reconocimiento.

Ahora bien, actualmente, con el desarrollo de los terminales multimedia, se generaliza cada vez más la captura de sonido denominada de "manos libres" con el fin de permitir al usuario escuchar mensajes vocales y leer simultáneamente informaciones presentadas sobre la pantalla de su terminal. En estas condiciones, el reconocimiento automático de la voz se convierte en más delicado de implementar, disminuyendo el nivel de la señal vocal útil mientras que el ruido ambiente permanece constante.

El documento JP 2003/241794 A divulga un procedimiento de sincronización entre por un lado una operación de procesamiento por reconocimiento automático de la voz de una secuencia vocal emitida por un locutor y, por otra parte, al menos una acción del locutor destinada a activar dicho procesamiento.

Por otra parte, como el usuario dispone hoy en día de otros medios distintos a la voz, se hace difícil para el sistema de reconocimiento determinar únicamente por la estructura del diálogo los instantes en los que el locutor va a pronunciar una secuencia vocal.

Para solucionar estos inconvenientes es por lo que ciertos terminales están equipados con un medio que permite al usuario activar el procesamiento de reconocimiento vocal sobre una acción ejercida, por ejemplo, sobre una tecla de un dispositivo conocido bajo el nombre de "Push-to Talk". Cuando el locutor inicia la emisión de una secuencia vocal, pulsa sobre la tecla de este dispositivo para indicar al servidor que la señal sonora que sigue es claramente una secuencia vocal que debe ser procesada por el sistema de reconocimiento de la voz. La tecla se libera por el locutor al final del enunciado de dicha secuencia vocal. De este modo, el sistema no busca reconocer al usuario más que cuando este pulsa sobre la tecla del dispositivo "Push-to Talk", evitando de este modo todas las falsas alertas fuera de los períodos en los que la tecla está pulsada.

Sin embargo, el dispositivo "Push-to Talk" presenta el inconveniente de que si el usuario comienza a emitir una secuencia vocal antes de pulsar sobre la tecla del dispositivo o si dicha secuencia se prosigue después de que el usuario haya liberado dicha tecla, el sistema de reconocimiento utilizará, no la secuencia real, sino una secuencia temporalmente truncada, lo que genera muchos errores de reconocimiento.

También, un problema técnico a resolver por el objeto de la presente invención es proponer, según la reivindicación 1, un procedimiento de sincronización entre, por una parte, una operación de procesamiento por reconocimiento automático de la voz de una secuencia vocal emitida por un locutor y, por otra parte, al menos una acción del locutor destinada a activar el procesamiento, lo que permitirá reducir los errores de reconocimiento que podrían sobrevenir debido a una sincronización imperfecta entre la acción de activación ejercitada por el locutor y el comienzo, e igualmente el final, de la secuencia vocal que pronuncia.

La solución al problema técnico planteado consiste, según la presente invención, en que la operación de procesamiento se efectúe a partir un instante dado t_0 precedente a la acción del locutor.

De este modo, se obtiene que, si el usuario ha activado tardíamente el procesamiento de reconocimiento con relación al comienzo de la secuencia vocal, pueden tenerse en cuenta por el sistema de reconocimiento unas informaciones temporalmente anteriores a su acción, pero relativas a la secuencia, con el fin de reducir los errores que se deberían a este defecto de sincronización.

Según un primer modo de realización, la operación de procesamiento consiste en una transferencia, a partir del instante dado t_0, de segmentos de voz extraídos de la secuencia vocal a un sistema de reconocimiento automático de la voz.

En este modo de realización, el procesamiento de la secuencia vocal por el sistema de reconocimiento de la voz se activa efectivamente en el momento de la acción del usuario, sobre la tecla "Push-to-Talk" por ejemplo, pero con un efecto retroactivo en el pasado, siendo transmitidos inmediatamente los segmentos de secuencia comprendidos en el intervalo de tiempo que separa el instante dado t_0 y la acción del usuario, al sistema de reconocimiento.

Como estos segmentos de secuencia pasados se transmiten sin retardo al sistema y el procesamiento del reconocimiento es en general más rápido que el tiempo real, el resultado del reconocimiento podrá suministrarse sin un retardo efectivo a pesar del tiempo necesario para que el sistema trate la parte de la secuencia vocal anterior a la acción del usuario.

La duración del intervalo de tiempo que precede a la acción del usuario para la toma en consideración de la parte de la secuencia vocal suplementaria a procesar debe elegirse suficientemente corta para no introducir un retardo demasiado grande, y suficientemente larga para permitir una compensación real del defecto de sincronización. A título de ejemplo, esta duración puede ser del orden de algunos centenares de milisegundos.

En la práctica, la invención prevé que dicha transferencia se efectúe a través de una línea de retardo. En particular, la línea de retardo se realiza mediante un registro circular.

Según un segundo modo de realización, la operación de procesamiento consiste en una validación, sobre la detección del comienzo de la actividad vocal entre el instante dado t_0 y un instante dado t_1, del reconocimiento automático de la voz efectuado sobre dicha secuencia vocal.

En este modo de realización, el procesamiento de reconocimiento de la voz se efectúa continuamente sobre el conjunto de las secuencias recibidas por el sistema, independientemente de cualquier acción del usuario sobre un dispositivo "Push-to-Talk" por ejemplo. En el transcurso de estas secuencias, cada detección de la voz genera un evento denominado de comienzo de actividad vocal. Estos eventos se validan o rechazan a continuación en función de su compatibilidad temporal con las acciones del usuario. Más precisamente, se validarán si se producen en un intervalo de tiempo que enmarca la acción del usuario, y se rechazarán fuera de este intervalo, el cual se determina por una duración de aceptación antes de la acción del usuario y una duración de aceptación después de la acción del usuario.

Se observará que el primer modo de realización de procesamiento retroactivo no aumenta la carga de cálculo del sistema de reconocimiento, pero puede introducir un retardo en la reacción del sistema si este no es suficientemente rápido para recuperar el retardo a partir de la acción del usuario sobre el dispositivo "Push-to-Talk". Por el contrario, el modo de realización en continuo no introduce ningún retardo, pero aumenta la carga de cálculo debido al procesamiento de las detecciones no compatibles con las acciones sobre el dispositivo "Push-to-Talk".

Los modos de realización que anteceden se refieran al problema de la sincronización de una acción del usuario con el inicio de la secuencia vocal.

Con referencia ahora a la sincronización de una acción del usuario con el final de la secuencia vocal, se prevé por la invención que el procesamiento de dicha secuencia vocal por el reconocimiento automático de la voz se prolongue más allá de una segunda acción del locutor que indica un final de la secuencia vocal.

El procedimiento de reconocimiento puede prolongarse en una duración que corresponde a la tolerancia admitida entre el final de la secuencia vocal y la acción de liberación de la tecla del dispositivo "Push-to-Talk" por ejemplo.

Es necesario observar que la indicación de fin de secuencia por una acción determinista del usuario es facultativa, siendo adecuado entonces el sistema de reconocimiento para detectar por sí mismo el final de dicha secuencia. En este contexto, se puede concebir que la acción de indicación del comienzo de la secuencia se traduzca por un simple impulso transmitido al sistema de reconocimiento automático de la voz. Cuando sea necesario, un nuevo impulso podrá indicar al sistema el final de la secuencia.

La descripción que sigue con relación a los dibujos adjuntos, dados a título de ejemplos no limitativos, permitirá comprender correctamente en qué consiste la invención y cómo puede realizarse.

La figura 1 es un cronograma que representa: (a) un registro sonoro efectuado mediante un micrófono de un terminal en comunicación con un sistema de reconocimiento automático de la voz, y (b) los elementos de dicho registro tenidos en cuenta para el procesamiento por el sistema de reconocimiento según un primer modo de realización del procedimiento de acuerdo con la invención.

La figura 2 es un esquema de un terminal para la implementación de dicho primer modo de realización de la invención.

La figura 3 es un esquema de un registro circular para la implementación de dicho primer modo de realización de la invención.

La figura 4 es un cronograma que representa: (a) un registro sonoro efectuado mediante un micrófono de un terminal en comunicación con un sistema de reconocimiento automático de la voz, y (b) los elementos de dicho registro tenidos en cuenta para el procesamiento por el sistema de reconocimiento según un segundo modo de realización del procedimiento de acuerdo con la invención.

En la figura 2 se representa un terminal equipado con un micrófono 11 principalmente destinado a registrar secuencias vocales emitidas por un locutor, principalmente durante sus intercambios con un servidor vocal 20 a través del sistema de reconocimiento automático de la voz. Obsérvese que el sistema de reconocimiento puede instalarse en un servidor o en el terminal en sí mismo. Sin embargo la invención es independiente de la implantación efectiva de este sistema.

Como lo muestra la figura 2, el terminal incluye igualmente un dispositivo "Push-to-Talk" 12 que permite, mediante una acción del usuario sobre el botón de este dispositivo, enviar al sistema de reconocimiento una indicación de inicio de secuencia vocal, así como eventualmente una indicación de final de secuencia.

La invención tiene por objeto solucionar los efectos negativos sobre el procesamiento del reconocimiento que puede tener una mala sincronización entre las acciones del usuario y el comienzo o el final de la secuencia vocal pronunciada.

La figura 1(a) muestra un registro bruto de sonidos captados por el micrófono 11 del terminal. Este registro hace aparecer una primera secuencia S_b correspondiente por ejemplo al ruido ambiente. Después, comienza una secuencia vocal S_v emitida por el locutor durante su diálogo con el servidor 20. Con el fin de informar a dicho servidor de que la secuencia S_v que pronuncia debe ser procesada por el sistema de reconocimiento, el usuario acciona en el instante t_d una tecla del dispositivo "Push-to-Talk" 12 que activa o bien un impulso, o bien la apertura de una ventana, como lo muestra la figura 1(b).

El procesamiento del reconocimiento de la secuencia S_v se hace efectivo inmediatamente desde el instante t_d. Puede sin embargo que el usuario haya sido tardío en su acción y que el comienzo de la secuencia S_v indicado en trazo de puntos se escape del procesamiento de reconocimiento, con todas las consecuencias nefastas que este truncado puede tener sobre la calidad final del reconocimiento.

Para resolver este problema, la invención prevé tener en cuenta en el procesamiento del reconocimiento a efectuar unos segmentos de voz contenidos en la parte truncada de la secuencia vocal S_v a partir del instante t_0 anterior al instante t_d de la acción del usuario.

En el ejemplo de la figura 1, el intervalo de tiempo δt entre los instantes t_0 y t_d es suficientemente largo para restaurar completamente la parte faltante de la secuencia S_v. En este caso, no se alterará la calidad final del reconocimiento. Por el contrario, si el retardo es demasiado grande, la parte truncada de la secuencia no se tomará en consideración más que parcialmente, lo que conducirá sin embargo a una mejora del reconocimiento, pero inferior a la que se obtendría con una reconstitución total.

Por supuesto, el intervalo de tiempo δt no debe ser demasiado largo para no introducir un desfase en el reconocimiento que fuera sensible al usuario. No debe ser demasiado corto, porque si no haría ineficaz el procedimiento de sincronización, objeto de la invención.

La implementación de este procedimiento de sincronización utiliza una línea 13 de retardo dispuesta en serie sobre el micrófono 11, tal como lo muestra la figura 2.

Dicha línea 13 de retardo puede realizarse mediante un registro circular, como se indica en la figura 3. Las muestras de la secuencia vocal numeradas 1, 2, 3,… se escriben, y posteriormente se leen y transfieren al sistema de reconocimiento, con la misma velocidad, pero con un retardo en número de muestras correspondiente al retardo temporal δt fijado. Después de haber sido leídas, las muestras se destruyen progresivamente.

Es necesario precisar por otro lado que la línea de retardo podría actuar igualmente sobre las tramas de análisis de la señal que se envían al sistema de reconocimiento. Más generalmente, la línea de retardo puede actuar en cualquier entorno del flujo entre la señal de micrófono y la entrada del sistema de reconocimiento.

Como ya se ha mencionado, es posible transferir en bloque al sistema de reconocimiento la parte restaurada de la secuencia, lo que limita un eventual retardo en el reconocimiento, esto mientras que el procedimiento de reconocimiento sea más rápido que el tiempo real.

La ventaja de este primer modo de realización que utiliza una línea de retardo es que no aumenta la carga de cálculo del sistema.

La figura 4 ilustra otra implementación de la invención que consiste en tratar continuamente la señal 4(a) procedente del micrófono 11 del terminal mediante un detector de actividad vocal. El resultado de este procesamiento se muestra en 4(b) con la toma en consideración de la secuencia S_v después de que se haya detectado como secuencia vocal en el instante t_a. Incluso si la acción del usuario sobre el dispositivo "Push-to-Talk" surge en el instante t_d posterior al instante t_a de detección de la actividad vocal, la secuencia se validará sin embargo si esta detección tiene lugar durante el intervalo $\delta t=\delta t_0+\delta t_1$ entre los instantes t_0 y t_1, Como se ve en la figura 4, δt_0 es una duración de aceptación antes de la acción del usuario mientras que δt_1 es una duración de aceptación después de la acción del usuario.

Este modo de realización no introduce ningún retardo porque actúa únicamente por filtrado sobre el procesamiento normal de reconocimiento, pero, por el contrario, aumenta la carga de cálculo del sistema.

Como lo muestra la figura 1, el usuario puede igualmente indicar el final de la secuencia vocal mediante una segunda acción sobre el dispositivo "Push-to Talk" en un instante t_f que genera o bien un impulso de final de secuencia, o bien el cierre de la ventana abierta al comienzo de la secuencia. Puede suceder también en este caso que el usuario haya actuado demasiado pronto, es decir antes del final de la secuencia vocal S_v. En lugar de cortar bruscamente la adquisición de la voz del locutor en el instante t_f, se propone prolongar esta adquisición y alimentar el sistema de reconocimiento durante una duración $\delta t'$ correspondiente a las tolerancias sobre la desincronización entre la acción del usuario y el final de la secuencia S_v. Es necesario observar que este procedimiento no introduce ningún retardo suplementario sino que reduce simplemente los truncados en el final del enunciado.

REIVINDICACIONES

1. Procedimiento de sincronización entre, por una parte, una operación de procesamiento por reconocimiento automático de la voz de una secuencia vocal (S_v) emitida por un locutor y, por otra parte, al menos una acción de dicho locutor destinada a activar dicho procesamiento, en que la operación de procesamiento se efectúa a partir un instante (t_0) precedente a dicha acción del locutor, siendo el intervalo de tiempo que separa dicho instante (t_0) y dicha acción del locutor un intervalo de tiempo dado.

2. Procedimiento según la reivindicación 1, caracterizado por que dicha operación de procesamiento consiste en una transferencia, a partir de dicho instante (t_0), de segmentos de voz extraídos de dicha secuencia vocal (S_v) a un sistema de reconocimiento automático de la voz.

3. Procedimiento según la reivindicación 2, caracterizado por que dicha transferencia se efectúa a través de una línea (13) de retardo.

4. Procedimiento según la reivindicación 3, caracterizado por que dicha línea (13) de retardo se realiza mediante un registro circular.

5. Procedimiento según la reivindicación 1, caracterizado por que dicha operación de procesamiento consiste en una validación, sobre la detección del comienzo de la actividad vocal entre dicho instante (t_0) y un instante dado (t_1), del reconocimiento automático de la voz efectuado sobre dicha secuencia vocal (S_v).

6. Procedimiento según una cualquiera de las reivindicaciones 1 a 5, caracterizado por que el procesamiento de dicha secuencia vocal (S_v) por el reconocimiento automático de la voz se prolonga más allá de una segunda acción del locutor que indica un final de la secuencia vocal (S_v).

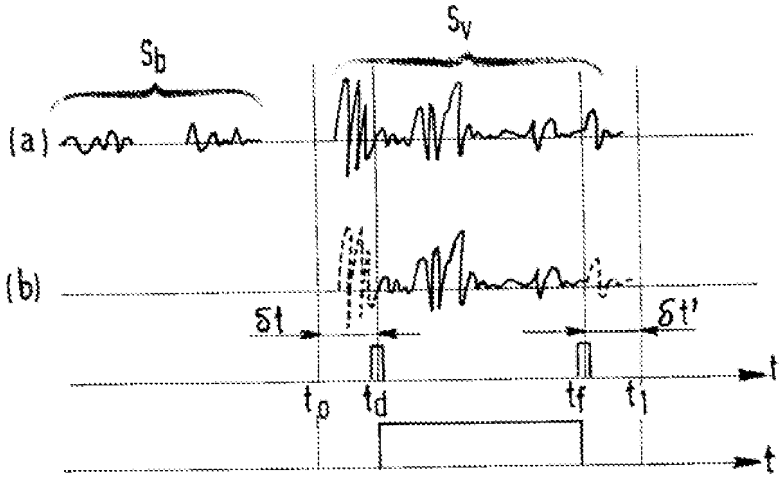

FIG.1

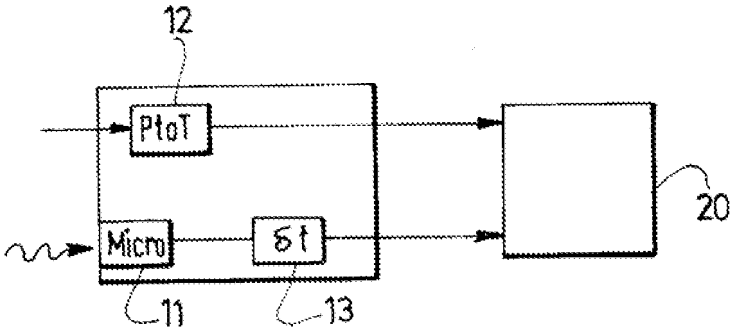

FIG.2

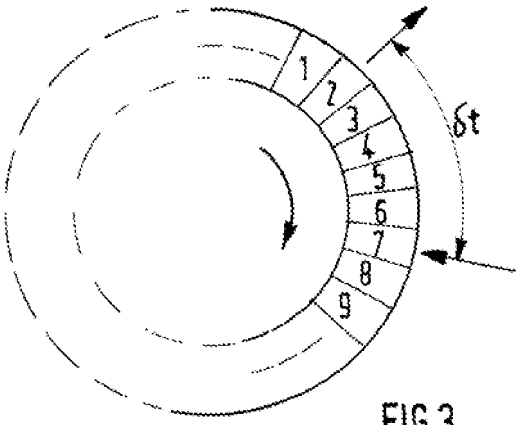

FIG.3

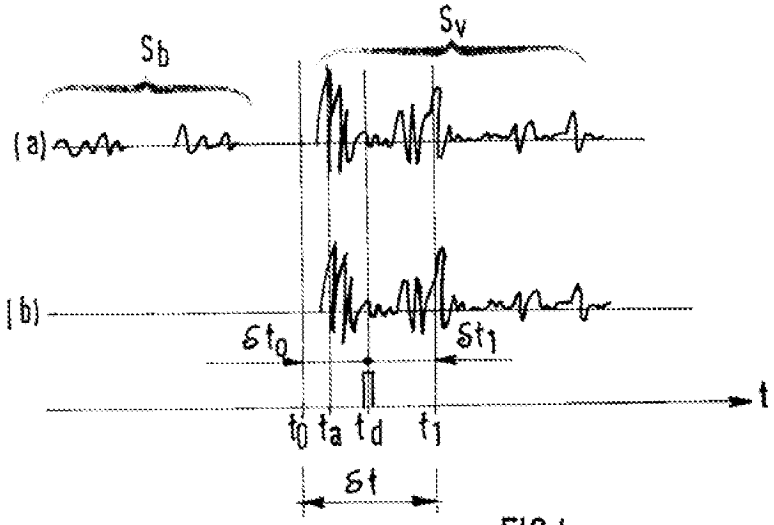

FIG.4