(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 369 665**

(51) Int. CI.:
G06F 17/27 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **03817059 .3**
(96) Fecha de presentación: **28.05.2003**
(97) Número de publicación de la solicitud: **1627325**
(97) Fecha de publicación de la solicitud: **22.02.2006**

(54) Título: **SEGMENTACIÓN AUTOMÁTICA DE TEXTOS QUE COMPRENDEN FRAGMENTOS SIN SEPARADORES.**

(45) Fecha de publicación de la mención BOPI:
02.12.2011

(45) Fecha de la publicación del folleto de la patente:
02.12.2011

(73) Titular/es:
LOQUENDO SPA
VIA ARRIGO OLIVETTI 6
10100 TORINO, IT

(72) Inventor/es:
BADINO, Leonardo

(74) Agente: **Ponti Sales, Adelaida**

DESCRIPCION

Campo de la invención

[0001] La invención se refiere a la segmentación de los textos en lenguajes que comprende fragmentos escritos sin separadores, como por ejemplo, espacios, guiones o similares. Ejemplo de un lenguaje como tal es el lenguaje chino mandarín, donde los fragmentos son normalmente representados por ideogramas.

[0002] Como es bien sabido por los expertos en técnica de la síntesis de voz, mediante "fragmentos" de un elemento de expresión se prevé que con mayor frecuencia corresponde a una palabra. Además del mandarín chino existen otros lenguajes en que, sin embargo, una sola palabra puede de hecho comprender varios fragmentos: un ejemplo típico de esto es el Alemán, donde existen palabras complejas, tales como "Patentübereinkommen" que, a pesar de que comprende dos bloques distintos, a saber "Patent" y " Übereinkommen" se escriben como una sola palabra sin separadores.

[0003] El resto de esta descripción sin embargo, se hará (sin que esto tenga que ser interpretado como una limitación del ámbito de aplicación de la invención) con referencia al chino mandarín, ya que este es uno de los lenguajes en que la invención puede ser aplicada de forma más beneficiosa.

Descripción de la técnica relacionada

[0004] La forma escrita del lenguaje representa una dificultad básica para los profanos que deseen aprender el lenguaje chino. De hecho, el conjunto de las "letras" para el chino incluye alrededor de 45 mil ideogramas ("hanzhi" en chino). Una buena parte de estos ideogramas son palabras (palabras compuestas de una sola letra) en relación con los objetos que ya no existen y que, por lo tanto, se han vuelto prácticamente inútiles. Una estimación actual es que con el fin de estar en condiciones de leer un periódico chino puede ser suficiente el conocimiento de cerca de 4.000 ideogramas.

[0005] Ya se trate de 4.000 o 40.000 ideogramas, el orden de magnitud es, en cualquier caso, mucho mayor que el conjunto de caracteres de los lenguajes indoeuropeos.

[0006] A partir de esto, surge una dificultad básica en el desarrollo de sistemas para la síntesis de texto a voz del chino. De hecho, para los lenguajes indoeuropeos la codificación de un solo carácter por medio de un dígito binario incluyendo ocho bits (es decir, un byte) de acuerdo con la norma ISO por lo general bastará. Por el contrario, para los chinos son necesarios por lo menos dos bytes para la codificación de cada ideograma individual.

[0007] La norma ISO no prevé este tipo de codificación, pero existen otras técnicas de codificación que lo pueden resolver, por ejemplo, como se demuestra por las técnicas de codificación conocidas como Unicode, GB y BIG5.

[0008] Recurrir al "pinyin" de alguna manera puede paliar el problema de la codificación. El pinyin es una forma de transcripción fonética/transliteración basada en caracteres latinos que muestran cómo se pronuncian las palabras en chino. La transcripción pinyin se proporciona en los libros de enseñanza de los fundamentos del lenguaje chino y en los diccionarios chinos y, como tal, es conocida para una buena cantidad de hablantes de chino.

[0009] Otra de las características básicas del lenguaje chino mandarín es que los ideogramas (es decir, los fragmentos de que se compone el lenguaje) se escriben sin separadores. En consecuencia, la identificación de cada palabra dentro de una frase no es nada fácil ya que cada palabra puede constar en realidad de uno o más hanzhis.

[0010] Uno puede ser llevado a creer erróneamente que este problema podría evitarse fácilmente con sólo la trascripción de un carácter (es decir, un ideograma) a la vez, sin preocuparse de donde termina una palabra determinada, y comienza una nueva.

[0011] En realidad, a fin de lograr una calidad aceptable en la síntesis de voz, es necesario que (incluso si los ideogramas se transcriben en forma pinyin) el texto debe ser descompuesto en palabras individuales.

[0012] Esta necesidad está determinada por una serie de factores,

- cada ideograma individual puede tener diferentes formas de pronunciación en función de las palabras a las que pertenece;

- ciertas reglas fonológicas y fonéticas dependen de la separación correcta de las palabras: por ejemplo, una regla fonológica llamada de tonos sandhi establece que en presencia de dos sílabas cada una transmite un tercer tono, la primera va a cambiar su tono, si las dos sílabas pertenecen a la misma palabra; y

- la información relativa a cada palabra es necesaria para permitir un correcto análisis gramatical y sintáctico-prosódico.

[0013] En resumen, un arreglo eficaz para segmentar el texto en partes es un requisito básico para una síntesis de voz verdaderamente satisfactoria de texto a voz del lenguaje chino mandarín.

[0014] La solución conocida para segmentar en fragmentos el texto en chino mandarín puede ser esencialmente dividida en tres categorías, a saber:

- algoritmos puramente estadísticos, como los llevados a cabo a través del llamado árbol de clasificación y regresión (CART),

- algoritmos basados en reglas léxicas, y

- algoritmos que combinan las dos soluciones anteriores.

[0015] Una primera aproximación (a veces conocida como la segmentación coincidente máxima o MMS), prevé que una frase se segmente en palabras sobre la base de un léxico dado mediante el intento de resolver lo mejor posible cualquier ambigüedad en relación con una frase dada estando adaptada a ser descompuesta en varias formas, extrayendo así palabras diferentes.

[0016] Para resolver esa ambigüedad, se utilizan con frecuencia las soluciones heurísticas como el criterio de máxima coincidencia posible perfeccionado además por otros criterios. La correspondencia máxima se basa en el reconocimiento del hecho de que, como regla general, la probabilidad de que una determinada secuencia de ideogramas pertenezca a una sola palabra en el léxico es más alta que la probabilidad de que dicha secuencia corresponda a una pluralidad de palabras más cortas concatenadas en el texto.

[0017] En las versiones más fáciles, el algoritmo busca, a partir del comienzo de la frase, y recurriendo a su propio léxico, la palabra compuesta por el mayor número de ideogramas. Después de localizar dicha palabra, el algoritmo analiza el ideograma inmediatamente próximo a la palabra acabada de encontrar y comienza la búsqueda de nuevo.

[0018] Los enfoques mixtos proporcionan un coste fijo que se asocia a cada palabra. Este coste se asigna siguiendo una métrica que puede estar relacionada con la frecuencia de aparición de la palabra en un lenguaje determinado o la probabilidad de que la categoría gramatical a la que la palabra pertenece pueda aparecer en el contexto sintáctico de la frase.

[0019] Entre los diferentes tipos de segmentaciones definido para una frase dada, la que tiene un coste mínimo es la seleccionada.

[0020] Los ejemplos de tales enfoques de la técnica anterior son, por ejemplo el artículo de R. Sproat et al. "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese", Computacional Linguístics, Volumen 22, Número 3, 1997 páginas 378-402 y US-A-6 173 252.

[0021] En concreto, el acuerdo descrito en el artículo de Sproat et al. prevé una función de costes que se aplica que es inversamente proporcional a la frecuencia de aparición de una determinada palabra en el vocabulario correspondiente.

[0022] Por el contrario, la disposición de US-A-6173252 es esencialmente del tipo basado en sintaxis, es decir, de la clase, donde las funciones de coste/peso son las relacionadas por ejemplo con las cadenas de error habitualmente cometidos, nombres de personas, lugares y organizaciones, números, y combinaciones de los números y las palabras como unidades de medida comunes a lo largo de la segmentación de palabras en un diccionario tradicional.

[0023] El artículo de Jiansheng Yu Yu y Shiewn "Some problems of Chinese segmentation", Institute of Computational Linguistics, Universidad de Pekín, Beijing, República Popular de China, 22 de marzo de 2003, analiza los principales problemas de la segmentación del chino, resumiendo los algoritmos existentes con las comparaciones teóricas, introduciendo de un léxico dinámica y discutiendo cómo el tamaño del léxico influye en la calidad de un proceso de segmentación.

[0024] El documento US-A-5.806.021 describe un segmentador automático que aplica dos procedimientos estadísticos para la segmentación de un texto continuo, un primer procedimiento de coincidencia adelante-atrás, adecuado para aplicaciones donde la velocidad es una preocupación, y un segundo procedimiento de búsqueda estadística de la pila, que es más preciso y requiere más tiempo de ejecución.

[0025] En el artículo de Jian-Yun Nie et al. "Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge", Communications of colips, Chinese and Oriental Languages Information Processing Society, SG, vol 5, no. 1-2, diciembre de 1995, páginas 47-57, se describe un proceso de segmentación basado en un enfoque híbrido haciendo uso tanto de la regla y el enfoque basado en diccionarios y el enfoque estadístico.

Objeto y descripción de la invención

[0026] Por tanto existe la necesidad de soluciones mejoradas para la síntesis de texto a voz de los lenguajes que (como el lenguaje chino mandarín) incluyen fragmentos sin separadores.

[0027] El objeto de la presente invención es proporcionar un arreglo mejorado.

[0028] De acuerdo con la presente invención, dicho objeto se logra por medio de un procedimiento con las características previstas en las reivindicaciones que siguen.

[0029] La invención se refiere también a un segmentador que opera de acuerdo con este procedimiento, estando el segmentador preferentemente en forma de un ordenador de propósito general adecuadamente programado. Por esta razón, la invención también se refiere a un producto de programa informático que se puede cargar en la memoria de un ordenador y que incluye porciones de código de software para realizar el procedimiento de la invención, cuando el producto se ejecuta en un ordenador. Además, la invención abarca un sistema de síntesis de texto a voz que incluye el segmentador mencionado en los párrafos anteriores.

[0030] Una característica importante de la invención radica en las diferentes métricas utilizadas en relación con la técnica anterior. En concreto, la invención tiene en cuenta el contexto semántico de cada palabra individual. De esta manera, la segmentación de una frase en un texto se hace depender de las frases anteriores (siempre que exista una correlación semántica) y el coste asignado a cada palabra varía en función de las palabras encontradas en las segmentaciones anteriores.

[0031] Todas las descomposiciones así obtenidas pueden por lo tanto ser asignadas en una red o matriz donde cada elemento se compone de una palabra más el coste correspondiente. Posteriormente, se elige la segmentación que tiene el coste más bajo, por ejemplo, mediante el uso de programación dinámica.

[0032] La invención se describirá ahora con referencia a las figuras adjuntas, en las que:

- Las figuras 1 a 4 están cada uno constituida por un diagrama de flujo que contiene una secuencia de etapas realizadas en el arreglo según la invención, y

- La figura 5 es un diagrama de bloques esquemático básico de un sistema correspondiente.

Descripción detallada de una realización preferida de la invención

[0033] A modo de introducción, se proporciona una descripción general de los principios básicos del arreglo descrito en este documento.

[0034] En resumen, el arreglo de la síntesis de texto a voz descritos en este documento se basa en un enfoque léxico sustancialmente relacionado con el enfoque de máxima coincidencia.

[0035] Como primer paso, el texto de entrada se subdivide en sintagmas siguiendo algunas reglas básicas, en que un sintagma es una parte del texto, por ejemplo, una frase, delimitada por signos de puntuación. A partir de entonces cada sintagma se envía a su vez al módulo de segmentación.

[0036] Más específicamente, a partir del primer ideograma (es decir, fragmento) en el sintagma, se buscan las secuencias "especiales" correspondientes a las reglas definidas (como fechas, horas, etc.). Si se ubican, a dichas secuencias se les asigna un coste definido.

[0037] Además, se busca la palabra más larga del léxico a partir de ese ideograma, luego la segunda más larga, y así sucesivamente terminando con el ideograma en sí mismo.

[0038] Aquellas palabras que se encuentran en el léxico tienen todas el mismo coste (por ejemplo, un coste igual a 5), mayor que el coste asignado a las secuencias especiales (por ejemplo coste igual a 3). Para aquellas palabras que no se ubican (es decir, se encuentran), ya sea mediante la búsqueda en base a normas o por la búsqueda en el léxico, se le asigna un coste más alto en relación con los costes considerados anteriormente.

[0039] De esta manera, se crea una especie de red o de matriz con tantas columnas como los ideogramas en el sintagma, por lo que un ideograma puede ser asociado a cada columna. El número de líneas varía en función de las columnas y se corresponde con el número de palabras incluidas en el léxico con el ideograma correspondiente a la columna como el primer ideograma.

[0040] Si no se encuentran palabras a partir de una columna dada, el número de líneas es fijo (con algunas excepciones), e incluye una palabra de longitud unitaria, entonces la palabra con el ideograma siguiente y así sucesivamente hasta una longitud dada.

[0041] A continuación se proporciona un ejemplo en el que en lugar de los ideogramas chinos se utilizan caracteres latinos, tales como A, B, C, D, etc. como representativos de los elementos individuales que comprenden los sintagmas sujeto a la segmentación.

[0042] Se supone que un léxico Lex está disponible, incluyendo un número de palabras imaginarias:

Lex = {A, ABC, BC, CD, CDAC, D}

y se considera la frase ABCDACEFD.

[0043] La red o matriz se organizará de la siguiente manera:

Columnas	0	1	2	3	4	5	6	7	8
líneas	ABC	BC	CDAC	D	A	CD	E	F	D
	A		CD				EF	FD	
							EFD		

[0044] En las posiciones designadas 6 y 7 no se encontraron palabras, por lo que a aquellas palabras en las columnas 6 y 7 se adjudica un coste que aumenta a medida que aumenta la longitud y que es más alto que el coste asignado a una palabra en el léxico con la misma longitud.

[0045] En este ejemplo, como es el caso en casi todas las frases del lenguaje chino, son posibles varias segmentaciones, por ejemplo ABC-D-A-EF-D o AB-CDAC-F-F-D.

[0046] El arreglo aquí descrito busca la secuencia con el menor coste. Esto se hace preferiblemente por medio de la programación dinámica, que puede ser fácilmente recurrida una vez que la red o la matriz se han creado. La programación dinámica conduce a un ahorro sustancial en términos de cálculos en comparación con aproximaciones de "fuerza bruta", donde se determinan todas las posibles secuencias y los costes respectivos.

[0047] A partir de la última posición en la frase/sintagma (por ejemplo, la posición 8) de la secuencia con el menor coste se busca para cada palabra en la columna. Con referencia a lo anterior esto es, al principio, D.

[0048] Una palabra dada identificada por la línea j y la columna i (en adelante denominado simplemente como Wi,j) la secuencia de menor coste a partir de Wi,j está dada por la siguiente fórmula:

$$MincostW_{i,j}=Min_{(k)}\{CostW_{i,j}+MinCostW_{(i+lengthWi,j),k}\}$$

[0049] Ciertamente pueden existir situaciones donde, a partir de la palabra Wi,j, existen varias secuencias posibles con el mismo coste, especialmente si la palabra está al final del sintagma.

[0050] En tal situación, por lo menos dos procedimientos heurísticos se pueden utilizar para la selección de una secuencia. Una primera opción es seleccionar la secuencia con la primera palabra más larga. Una alternativa es seleccionar la secuencia que tiene la variación de longitud inferior.

[0051] A modo de explicación, el arreglo que acabamos de describir se comparará ahora con una solución que operan sobre la base de un enfoque (puramente) léxico.

[0052] A modo de ejemplo, la frase/sintagma ABCDAC se tendrá en cuenta para la segmentación utilizando un enfoque de máxima coincidencia al referirse al mismo léxico considerado en el anterior.

[0053] La secuencia en cuestión puede de hecho ser segmentada solamente de una única manera, es decir, AB-CDAC. Sin embargo, una solución de coincidencia máxima generalmente localizaría la secuencia incompleta ABC-D-A y, posteriormente, se detendría sin haber encontrado la secuencia correcta.

[0054] Por supuesto, recurrir a un paso de retroceso podría prescindir de este inconveniente, pero esto supondría una carga importante en términos de complejidad computacional, que a su vez afectaría negativamente a lo que es considerado actualmente como el punto fuerte de un acercamiento de coincidencia máxima.

[0055] El arreglo conocido como MMS es esencialmente un algoritmo básico que hace uso del concepto heurístico de coincidencia máxima.

[0056] Los ejemplos de este enfoque es la solución conocida como MMSEG, para obtener información general acerca de MMSEG se puede hacer referencia por ejemplo, al artículo de Chih-Hao Tsai: " MMSEG: A Word Identification System for Mandarin Chinese Text Based On Two Variations of the Maximum Matching Algorithm", CHIN-HAO TSAI'S TECHNOLOGY PAGE, [en línea] 12 de marzo de 2000, XP002269269 Obtenido de Internet: <URL:http://technology.chtsai.org/mmseg/>.

[0057] MMSEG es sin duda uno de los segmentadores más eficaces que hacen uso del concepto de máxima coincidencia. Sin embargo, como el MMS (aunque con una probabilidad mucho más baja), puede no encontrar una secuencia correcta a pesar de que exista. También en este caso dar marcha atrás puede representar una solución a ese problema.

[0058] En concreto, MMSEG elige, a partir del inicio del sintagma, la primera palabra en una secuencia de tres fragmentos que tiene la longitud máxima. Por ejemplo, asumiendo un léxico Lex = {A, B, AB, CD, E, EF} y la frase ABCDEFABCD, MMSEG busca todas las posibles secuencias comprendidas en una ventana de tres fragmentos, es decir:

(1) A-B–CD

(2) AB-CD-E

(3) AB-CD-EF

[0059] A continuación, se selecciona la primera palabra de la secuencia más larga (la secuencia 3) que corresponde a AB.

[0060] En consecuencia, MMSEG logra buenos resultados. Sin embargo, además de tener una carga computacional apreciable, tiene la limitación de no tener en cuenta todas las posibles secuencias con el consiguiente riesgo de no aplicar de manera coherente el criterio heurístico de coincidencia máxima.

[0061] A modo de ejemplo adicional, se puede hacer referencia a la Lex léxico = (A, AB, BC, CD, DE, EF, GH, I, FGHI) y la frase ABCDEFGHI. El arreglo MMSEG no será capaz de localizar la palabra FGHI a pesar de que esta está incluida en una secuencia aceptable (A-CD-DE-FGHI).

[0062] El arreglo aquí descrito prescinde de este inconveniente, ya que puede tener en cuenta todas las posibles secuencias sin excluir ninguna secuencia. De esta manera se evita el riesgo de no detectar las palabras que tienen una alta probabilidad de ser las correctas de acuerdo con el criterio de máxima coincidencia.

[0063] Los llamados algoritmos estadísticos difieren ligeramente de los algoritmos con una base léxica debido a su mejor comportamiento en la segmentación de palabras desconocidas (es decir palabras que no están incluidas en el corpus de entrenamiento), tales como los nombres personales. El arreglo según la invención sufre parcialmente la misma desventaja, pero puede complementarse con normas que hacen que sea más fácil reconocer los símbolos específicos (por ejemplo: fechas, horas, etc.).

[0064] Una vez más hay que recordar aquí que una característica importante de los arreglos descritos en este documento se encuentra en las métricas diferentes utilizadas con respecto a la técnica anterior.

[0065] En concreto, el arreglo descrito en este documento tiene en cuenta el contexto semántico de cada palabra. La segmentación de una frase en un texto se hace depender por lo tanto de las frases anteriores (siempre existe

una correlación semántica) y el coste asignado a cada palabra varía en función de las palabras encontradas en las segmentaciones anteriores.

[0066] Todas las descomposiciones así obtenidas se pueden asignar por lo tanto en una red o matriz en la que cada elemento se compone de una palabra más el coste correspondiente. Posteriormente, se elige la segmentación que tenga el coste más bajo, por ejemplo, mediante el uso de programación dinámica.

[0067] Volviendo ahora a los diagramas de flujo de las figuras 1 a 4, se presumirá que el segmentador descrito en este documento acepta como entrada un texto codificado con el sistema Unicode (o un sistema similar), estando dicho texto subdividido en apartados, que a su vez se subdividen en "sintagmas" que es el texto delimitado en cadenas mediante secuencias de caracteres específicos (por ejemplo, un punto o una coma seguida de un espacio o una nueva línea, un signo de exclamación o de interrogación, un espacio en blanco entre dos ideogramas, y así sucesivamente).

[0068] En la figura 1, una etapa 100 designa en general la etapa correspondiente al texto entrada que se está entrando en el sistema, mientras que la etapa 110 es una etapa en que se hace una verificación de que el texto en cuestión no es nulo. Si este es el caso, entonces el proceso termina en una etapa 160.

[0069] De lo contrario, un párrafo es extraído del texto y se cargan en una memoria intermedia A (Figura 5). Esto ocurre en una etapa designada como 120.

[0070] En una etapa 130, la memoria intermedia A se comprueba para determinar si está vacía.

[0071] Si la memoria intermedia A no está vacía, se extrae un sintagma y se inserta en la memoria intermedia B. Esto ocurre en una etapa 140 después de que el sistema evoluciona de nuevo antes de la etapa 110.

[0072] Si la memoria intermedia A está vacía, el sistema evoluciona a una etapa 150 y luego de vuelta aguas arriba de la etapa 130.

[0073] Que el sistema regrese a la etapa 130 una vez que el sintagma se inserta en el memoria intermedia B significa que se proporciona la etapa 140, siendo la etapa 140 una etapa de espera para asegurarse de que todos los sintagmas en la memoria intermedia B han sido procesados por el segmentador para volver a la etapa 110 después de vaciar la memoria intermedia B.

[0074] Los expertos en la materia de inmediato se darán cuenta de que la subdivisión del texto en párrafos no es estrictamente necesaria. De hecho, el texto de entrada en su totalidad se puede considerar como un solo párrafo.

[0075] Una vez que la memoria intermedia B se ha rellenado con los sintagmas del párrafo actual, se extrae cada sintagma individual en una etapa 200 (figura 2) después de lo cual, en una etapa 210, la memoria intermedia B se comprueba para determinar si está vacía. Si este es el caso el léxico dinámico (ver más abajo) se vacía en una etapa 220 a evolucionar a la etapa 160. Si la etapa 210 obtiene un resultado negativo, el sistema evoluciona a la descomposición adecuada en palabras como lo demuestra la etapa 230.

[0076] La entrada del diagrama de flujo de la figura 3 es el único sintagma, designado como 300. En una etapa 304, un puntero (INDX) se ajusta en el primer carácter del sintagma (puntero ajustado a 0).

[0077] En una etapa 308 se busca la cadena más larga posible a partir del ideograma en la posición designada por el puntero INDX.

[0078] En este tipo de búsqueda, se buscan los llamados fragmentos "especiales": esto incluye, por ejemplo fechas, horas, números – tanto como ideogramas y como carácter latino - y las secuencias de caracteres diferentes de los ideogramas.

[0079] Si la etapa 312 indica un resultado positivo para la búsqueda, un fragmento nuevo se agrega a una memoria intermedia C (véase el gráfico 5) que tiene asociado su correspondiente coste fijo CF. Esto ocurre en una etapa designada como 316.

[0080] Por el contrario, si la búsqueda tuvo un resultado negativo (resultados negativos de la etapa 312) el sistema evoluciona directamente a la etapa 320, donde se lleva a cabo una nueva búsqueda.

[0081] En esta fase de la cadena se extrae del texto comprendido entre el ideograma en la posición indicada por el puntero INDX hasta un ideograma determinado (por ejemplo, el onceavo ideograma), si este no es el último ideograma en el sintagma. Si ocurre lo contrario, la cadena es la que existe entre INDX hasta el final del sintagma.

[0082] La cadena así obtenida se buscó entre las palabras incluidas en un léxico estático.

[0083] Si la búsqueda arroja un resultado positivo, la palabra localizada se escribe en la memoria intermedia C junto con el respectivo coste que equivale a un valor CM designado constante (que es generalmente mayor que CF). Posteriormente, la cadena se acorta al eliminar el último ideograma a la derecha y la búsqueda se repite.

[0084] Una vez que esta búsqueda se ha completado, la memoria intermedia C se actualiza mediante la inserción de todas las palabras ubicadas en el junto con sus costes, es decir, CM. Esto ocurre en una etapa designada como 324.

[0085] Posteriormente, en una etapa 328, si por lo menos una de las dos búsquedas ha dado un resultado positivo, el sistema evoluciona hacia una etapa designada como 332. Por otra parte, el sistema evoluciona directamente hacia una etapa 344. El coste de cada palabra presente en la memoria intermedia C se actualiza con el coste correspondiente en SLEX si la palabra está presente en SLEX y si su longitud es de al menos dos caracteres.

[0086] En la etapa 332, se actualizan los valores para el número de palabras localizadas ya presentes en sintagmas anteriores (NOL), más el recuento de todas las palabras ya localizadas (NW).

[0087] Una etapa designada como 336 corresponde a la actualización de un diccionario dinámico (SLEX), que será mejor detallado a continuación al referirse al diagrama de flujo de la figura 4.

[0088] Posteriormente, en una etapa 340, si ninguna búsqueda ha dado resultado, la palabra compuesta por el único ideograma en la posición designada por el puntero INDX se carga en el memoria intermedia C con un coste CS que es superior a la CM. Todavía en la etapa 340 todas las palabras de la memoria intermedia C se transfieren a una red o matriz RET (esto corresponde a las tablas previamente informadas en la descripción) en la columna designada por el puntero INDX.

[0089] A partir de entonces, en la etapa 344, el puntero INDX se incrementa en 1 y se realiza una comprobación en una etapa 348 si el valor resultante supera el último ideograma en el sintagma.

[0090] Si este no es el caso, tiene lugar una actualización dinámica del léxico SLEX en el que todos los costes de todas y cada entrada se incrementan en un valor constante, mientras que prescindiendo de los fragmentos que tienen un coste superior a CM. Esta tendrá lugar en una etapa designado 352.

[0091] Por el contrario, si el valor actualizado de INDX excede el último ideograma en el sintagma, en una etapa 356, el léxico dinámico es sometido a una actualización, mientras que los valores de NOL, NW y INDX se ponen a cero. En ese momento, el sistema evoluciona a la etapa 200.

[0092] El diagrama de la figura 4 detalla el proceso de actualización del léxico dinámico SLEX.

[0093] Cada palabra contenida en el memoria intermedia C (que se encuentra en la etapa designada como 400) se busca, en una etapa 410, en el léxico dinámico que está completamente vacío, cuando un nuevo párrafo comienza a procesarse (etapa 420).

[0094] Si la palabra ya estaba presente en el léxico dinámico, el coste relativo se reduce en un valor constante de CC en una etapa 430. Si la palabra no estaba presente en el léxico dinámico, se realiza una comprobación en una etapa 440 si el léxico dinámico está lleno.

[0095] Si este no es el caso, en una etapa 450 la palabra se inserta junto con el coste relativo (CM o CF) reducido en un valor de DCI.

[0096] Por el contrario, si el léxico SLEX dinámico está lleno, se realiza una comprobación en una etapa 460 si existen palabras que tengan un coste superior a CM.

[0097] Si este es el caso en una etapa 470, esa palabra es sustituida por la nueva palabra con un coste definido como en la etapa anterior 450.

[0098] Si no existe tal palabra que tenga un coste superior a CM, el sistema evoluciona directamente a la etapa 480. Se trata esencialmente de una comprobación a fin de determinar si todas las palabras en la memoria intermedia C han sido examinadas.

[0099] Si este no es el caso, el sistema evoluciona a la etapa 400. Si, por el contrario, todas las palabras en la memoria intermedia C se han examinado, el sistema evoluciona a una etapa final 490.

[0100] Se puede apreciar que el coste de cada palabra en el léxico dinámico nunca es menor de cero.

[0101] Una vez que la red o matriz RET ha sido completada, se ubicará la secuencia de coste mínimo. Preferentemente, se recurre a la programación dinámica para hacerlo.

[0102] En concreto, para cada palabra Wi,j en la red el coste mínimo se calcula para la secuencia de arranque de Wi,j sobre la base de la siguiente fórmula:

$$MincostW_{i,j}=Min(over\ k)\{CostW_{i,j}+MinCostW(i+lengthW_{i,j}),k\}$$

[0103] Dónde Mincost indica un coste mínimo, Min designa la función mínima (sobre k) y la longitud considerada es la longitud de la palabra Wi,j.

[0104] Si la palabra en cuestión contiene más de dos ideogramas, el factor de coste designado Cost Wi,j es una función de la proporción de NOL a NW que le da un sentido cuantitativo a la correlación semántica del sintagma actual con los sintagmas anteriores. Además, esta relación varía en función de si la palabra está ya presente en el léxico SLEX dinámico.

[0105] Preferiblemente, la función se define de la siguiente manera:

- Si la palabra no se había incluido previamente en el léxico dinámico, entonces

$$CostW_{i,j}=CSLEX$$

- De lo contrario

$$CostW_{i,j}=CSLEX+(Cfs-CSLEX)*(1-NOL/NW)/K$$

[0106] En las dos ecuaciones CSLEX representa el coste de la palabra en el léxico dinámico (SLEX), mientras que el Cfs es igual a CM o CF en función de si la palabra se encuentra por medio de la segunda búsqueda (B) o la primera búsqueda (A), mientras que K es un valor constante.

[0107] Estos costes se refieren a cada carácter.

[0108] Los expertos en la técnica de inmediato se darán cuenta de que los diagramas de flujo de las figuras 1 a 4 reflejan directamente en los correspondientes bloques funcionales de un respectivo segmentador 10 adaptado para ser aplicado sobre la base de la arquitectura que se muestra esquemáticamente en la figura 5, recurriendo a un ordenador tal como un procesador dedicado o un ordenador/procesador adecuadamente programado de propósito general o cualquier estructura equivalente de procesamiento de datos.

[0109] El segmentador 10 está a su vez adaptado a constituir un bloque de construcción básico de un sistema de síntesis de texto a voz que incluye un número de componentes de otros subsistemas generalmente designados como 30 y 40.

[0110] De estos subsistemas (que son de por sí conocidos en la técnica, por lo que no es necesario proporcionar una descripción detallada en este documento) el subsistema 30 incluye una instalación de introducción de texto, tal como un lector OCR, un teclado o cualquier otra fuente de texto adaptada para introducir texto como el texto de chino mandarín en el segmentador 10.

[0111] Dicha instalación para la introducción puede incluir (si no están incluidos en el segmentador 10) bloques de procesamiento – que no se muestran, pero son conocidos en la técnica - por ejemplo, adaptados para la codificación de los elementos individuales (es decir, los ideogramas) del que se compone el texto en cadenas de bits utilizando técnicas de codificación, como el estándar ISO, o las técnicas de codificación Unicode, GB o BIG5. La elección de la técnica de codificación, posiblemente, puede depender de los ideogramas que se sometieron a la transliteración fonética pinyin en vista de la segmentación en el segmentador 10.

[0112] La referencia 40 designa en su conjunto un subsistema de síntesis de voz - una vez más de tipo conocido por sí mismo - adaptado para transformar las secuencias resultantes de la segmentación en el segmentador 10 en la síntesis de los datos de emisión adaptados para generar una señal correspondiente de grabación de audio emitida por ejemplo, a través de un altavoz 50.

[0113] Por supuesto, sin perjuicio de los principios fundamentales de la invención, los detalles y las formas de realización pueden variar significativamente, con respecto a lo que se ha descrito a modo de ejemplo solamente, sin apartarse del alcance de la invención tal como se define mediante las reivindicaciones adjuntas.

REIVINDICACIONES

1. Procedimiento implementado por ordenador de segmentación en fragmentos, sintagmas de un texto escrito que incluyen elementos individuales, sin separadores, estando dichos fragmentos compuestos por cadenas incluyendo al menos uno de dichos elementos individuales, incluyendo el procedimiento las etapas de:

- Proporcionar un léxico que incluye un conjunto de cadenas, estando cada cadena compuesta de por lo menos uno de dichos elementos, en donde las cuerdas en dicho léxico son al menos parcialmente, representativas de dichos fragmentos, comprendiendo dicho léxico un léxico estático como un conjunto predeterminado de cadenas y un léxico dinámico,

- Buscar el sintagma que se segmenta sobre una base de elemento por elemento (INDX) mediante la búsqueda dentro de dicho léxico estático de cadenas correspondientes a cualquiera de dichos fragmentos, en el que, en el caso de un resultado positivo de búsqueda (312), el fragmento localizado correspondiente se almacena en una memoria intermedia (C) asociada a un coste correspondiente (CM),

- Comprobar si el fragmento localizado ya estaba presente en el léxico dinámico (SLEX) y:

a) en el caso de que el fragmento localizado ya estaba presente, reduciendo los costes asociados al mismo;

b) en el caso de que el fragmento localizado no existía previamente en el léxico dinámico, controlar (440) si el léxico dinámico está lleno y

i) si el léxico dinámico no está lleno, almacenar el fragmento localizado en el léxico dinámico con los costes respectivos (CM, CF) disminuidos en un valor constante (DCI),

ii) si el léxico dinámico está lleno, buscar cualquier fragmento almacenado que tenga un coste asociado mayor que un umbral de coste dado y, si se localizara dicho fragmento, sustituir el fragmento nuevo (450) por dicho fragmento;

- Almacenar, como resultado de dicha búsqueda, una pluralidad de secuencias de segmentación candidatas, cada una correspondiente a un modelo de segmentación respectivo y teniendo un coste devengado asociado correspondiente, y

- Seleccionar como el resultado final de la segmentación la secuencia candidata con el menor coste asociado acumulado.

2. Procedimiento según la reivindicación 1, caracterizado por el hecho de que, en presencia de dos secuencias candidatas que tengan el mismo coste asociado, se incluye la etapa de selección, como resultado de la segmentación de la secuencia candidata seleccionada del grupo que consiste en:

- La secuencia que tiene el primer fragmento más largo, y

- La secuencia que tiene la variación de longitud inferior.

3. Procedimiento según la reivindicación 1, caracterizado por el hecho de que al menos un sintagma en el texto de dicho ha sido previamente segmentado, caracterizado porque incluye las etapas de la determinación de al menos uno de:

- el número (NOL) de fragmentos situados en el sintagma instantáneo en que ya estaban presentes en dicho al menos un sintagma previamente segmentado, y

- la cantidad (NW) de fragmentos ya encontrados en el proceso de segmentación.

4. Procedimiento según la reivindicación 3, caracterizado por el hecho de que dicha secuencia que tiene el mínimo coste asociado es seleccionada sobre la base de una función de costes incluyendo al menos uno de dicho número de fragmentos (NOL), y dicha cantidad (NW).

5. Procedimiento según la reivindicación 3, caracterizado por el hecho de que dicha secuencia con el mínimo coste asociado es seleccionada sobre la base de una función de costes incluyendo la relación de dicho número de fragmentos (NOL), y dicha cantidad (NW).

6. Procedimiento según la reivindicación 1, caracterizado por el hecho de que incluye la etapa de aumento de dicho coste asociado (CM) por un valor constante en cada nueva etapa (INDX) en dicha búsqueda sobre una base de elemento por elemento.

7. Procedimiento según la reivindicación 6, caracterizado por el hecho de que incluye la etapa de prescindir de los fragmentos que tienen un coste mayor que un umbral dado (CM), cuando dicho coste asociado (CM) es mayor.

8. Procedimiento según la reivindicación 1, caracterizado por el hecho de que incluye, en el caso de un resultado positivo

de búsqueda (312), la etapa de reducir la cadena de búsqueda mediante la eliminación de uno de los elementos de sus extremos, repitiéndose entonces la búsqueda sobre la base de dicha cadena reducida.

9. Procedimiento según la reivindicación 8, caracterizado por el hecho de que incluye la etapa de reducción de dicha cadena, eliminando el elemento más a la derecha de la misma.

10. Procedimiento según la reivindicación 1, caracterizado por el hecho de que incluye las etapas de:

- definir al menos una parte de dicho conjunto de cadenas en dicho léxico (LEX) como representativas de fragmentos especiales que corresponden a reglas definidas,

- buscar el sintagma que está siendo segmentado sobre una base de elemento por elemento (INDX) mediante la búsqueda dentro de dicho léxico, al menos una de:

- (A) la cadena más larga correspondiente de cualquiera de dichos fragmentos especiales, en donde, en el caso de un resultado positivo de búsqueda (312), el fragmento correspondiente localizado se almacena en dicha memoria intermedia (C) con una primera capa asociada (CF),

- (B) la cadena más larga que corresponde a cualquiera de las otras cadenas en dicho léxico, en que, en el caso de un resultado positivo de búsqueda (324) el fragmento correspondiente localizado se almacena en dicha memoria intermedia (C) con un segundo coste asociado (CM), siendo dicho segundo coste (CM) mayor que dicho coste inicial (CF),

en el que si ninguna de dichas dos búsquedas (A) y (B) conducen a un resultado positivo, el elemento individual que se utilice como elemento de partida de la búsqueda se almacena en dicha memoria intermedia (C) con una tercera capa asociada (CS), siendo dicho tercer coste (CS) mayor que dicho segundo coste (CM).

11. Procedimiento según la reivindicación 10, caracterizado por el hecho de que incluye la etapa de aumento de dicho primer (CP), segundo (CM) y tercer (CS) coste mediante un valor constante en cada nueva etapa (INDX) en dicha al menos una búsqueda (A, B) sobre una base de elemento por elemento.

12. Procedimiento según la reivindicación 11, caracterizado por el hecho de que incluye la etapa de prescindir de los fragmentos que tienen un coste más alto que un umbral dado (CM), cuando dichos costes (CF, CM, CS) se incrementan.

13. Procedimiento según la reivindicación 12 caracterizado por el hecho de que dicho umbral determinado se selecciona igual a dicho segundo coste (CM).

14. Procedimiento según la reivindicación 10, en el que al menos un sintagma en dicho texto ha sido segmentado, caracterizado por el hecho de que incluye las etapas de:

- determinar el número (NOL) de fragmentos situado en el sintagma instantáeo en que ya estaban presentes en dicho al menos un sintagma previamente segmentado y el recuento (NW) de los fragmentos ya se encuentra en el proceso de segmentación,

- seleccionar dicha secuencia con el mínimo coste asociado es seleccionada sobre la base de una función de costes definida de la siguiente manera:

- i) si el fragmento localizado no se había incluido previamente en dicho léxico

$$CostW_{i,j}=CSLEX$$

- ii) en caso contrario

$$CostW_{i,j}=CSLEX+(Cfs-CSLEX)*(1-NOL/NW)/K$$

en el que el cfs es igual a dicho segundo coste (CM) o dicho coste inicial (CF), dependiendo de si la palabra considerada se localizó por medio de dicha segunda búsqueda (B) o de dicha primero búsqueda (A), K es un valor constante, CSLEX es el coste asociado al fragmento Wi,j en dicho léxico, y NOL y NO son dicho número y dicha cantidad, respectivamente.

15. Procedimiento según la reivindicación 1, caracterizado por el hecho de que incluye la etapa de codificación de dichos elementos individuales a cadenas de bits utilizando al menos uno del estándar ISO, o técnicas de codificación Unicode, GB o BIG5.

16. Procedimiento según la reivindicación 1, caracterizado por el hecho de que dichos elementos individuales se corresponden con los ideogramas.

17. Procedimiento según la reivindicación 16, caracterizado por el hecho de que dichos ideogramas son ideogramas del lenguaje chino mandarín.

18. Procedimiento según la reivindicación 17, caracterizado por el hecho de que incluye la etapa de transcribir dichos ideogramas en la transliteración fonética pinyin, antes que dichos sintagmas sean segmentados.

19. Procedimiento según la reivindicación 10, caracterizado por el hecho de que dichos fragmentos especiales son seleccionados del grupo formado por las fechas, horas y números.

20. Segmentador (10) para segmentar en fragmentos sintagmas de un texto escrito como elementos individuales, sin separadores, estando dichos fragmentos compuestos de cadenas incluyendo al menos uno de dichos elementos individuales, incluyendo el segmentador una estructura de procesamiento de datos (10, A, B, C, RET) configurada para llevar a cabo el procedimiento de cualquiera de las reivindicaciones 1 a 19.

21. Sistema de síntesis texto a voz (20), que incluye:

- Una fuente de texto (30) para generar al menos un sintagma de texto a ser segmentado en partes, incluyendo dicho sintagma elementos individuales por escrito sin separadores, estando dichos fragmentos compuestos de cadenas incluyendo al menos uno de dichos elementos individuales,

- Un segmentador (10) para recibir dicho al menos un sintagma de texto, incluyendo el segmentador una estructura de procesamiento de datos (10, A, B, C, RET) configurada para llevar a cabo el procedimiento de cualquiera de las reivindicaciones 1 a 19, generando así el resultado final de la segmentación teniendo dicha secuencia candidata el menor coste asociado, y

- Un generador de señal de voz (40, 50) para la conversión de dicha secuencia resultante de la segmentación en una señal de audio de voz correspondiente.

22. Producto de programa informático, que se puede cargar en la memoria de un ordenador e incluye porciones de código de software para llevar a cabo los pasos del procedimiento de cualquiera de las reivindicaciones 1 a 19.

Fig. 1

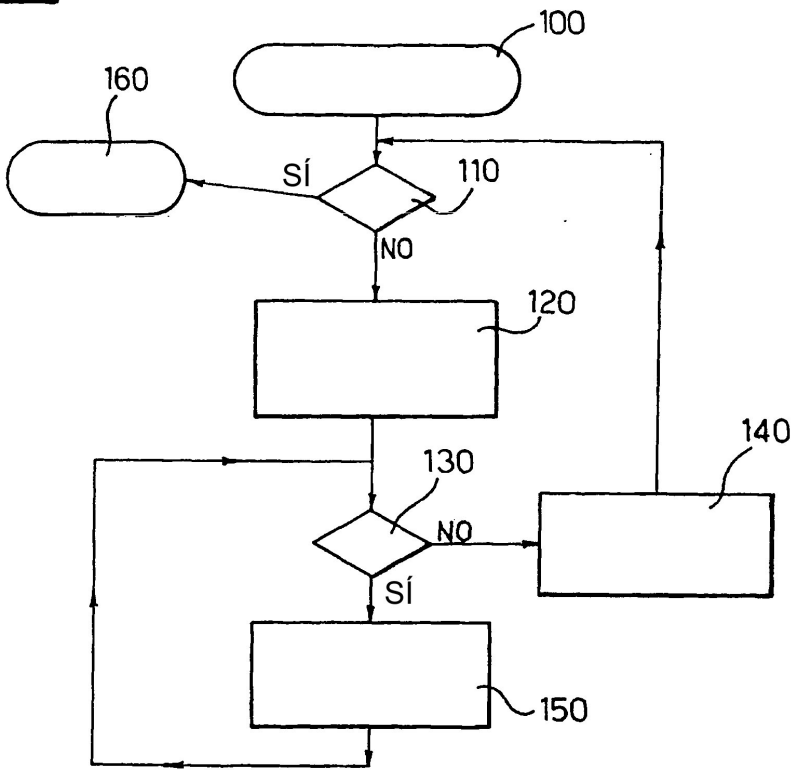


Fig. 2

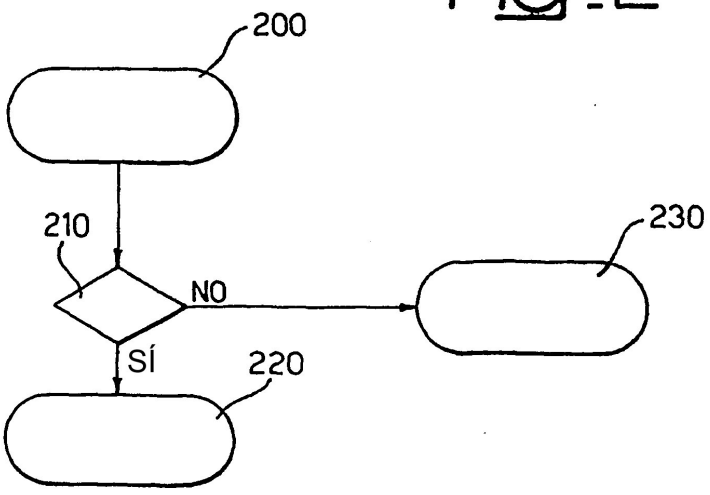


Fig.3

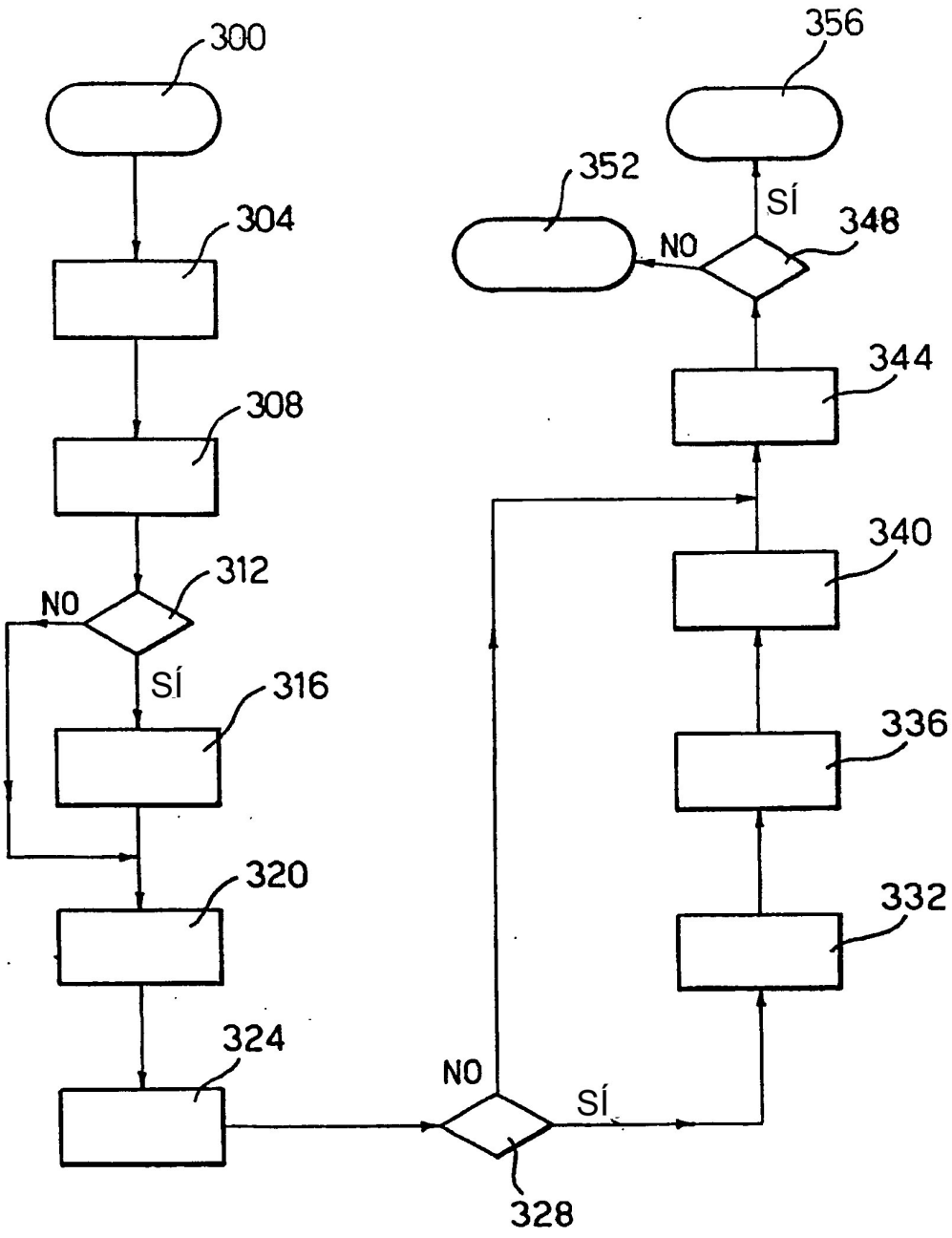
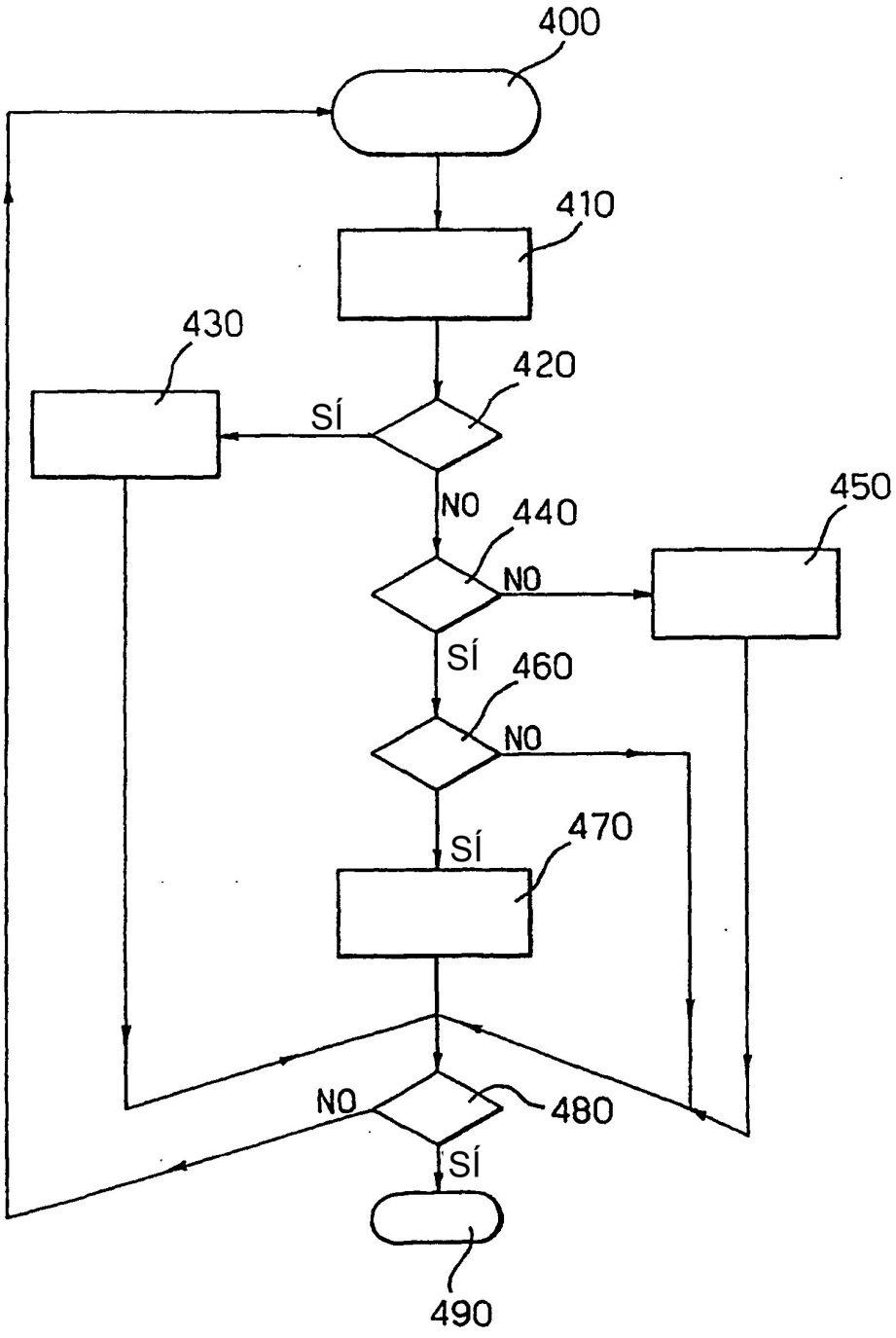

300
304
308
312
NO
SÍ
316
320
324
NO
SÍ
328
352
356
SÍ
NO
348
344
340
336
332

Fig. 4



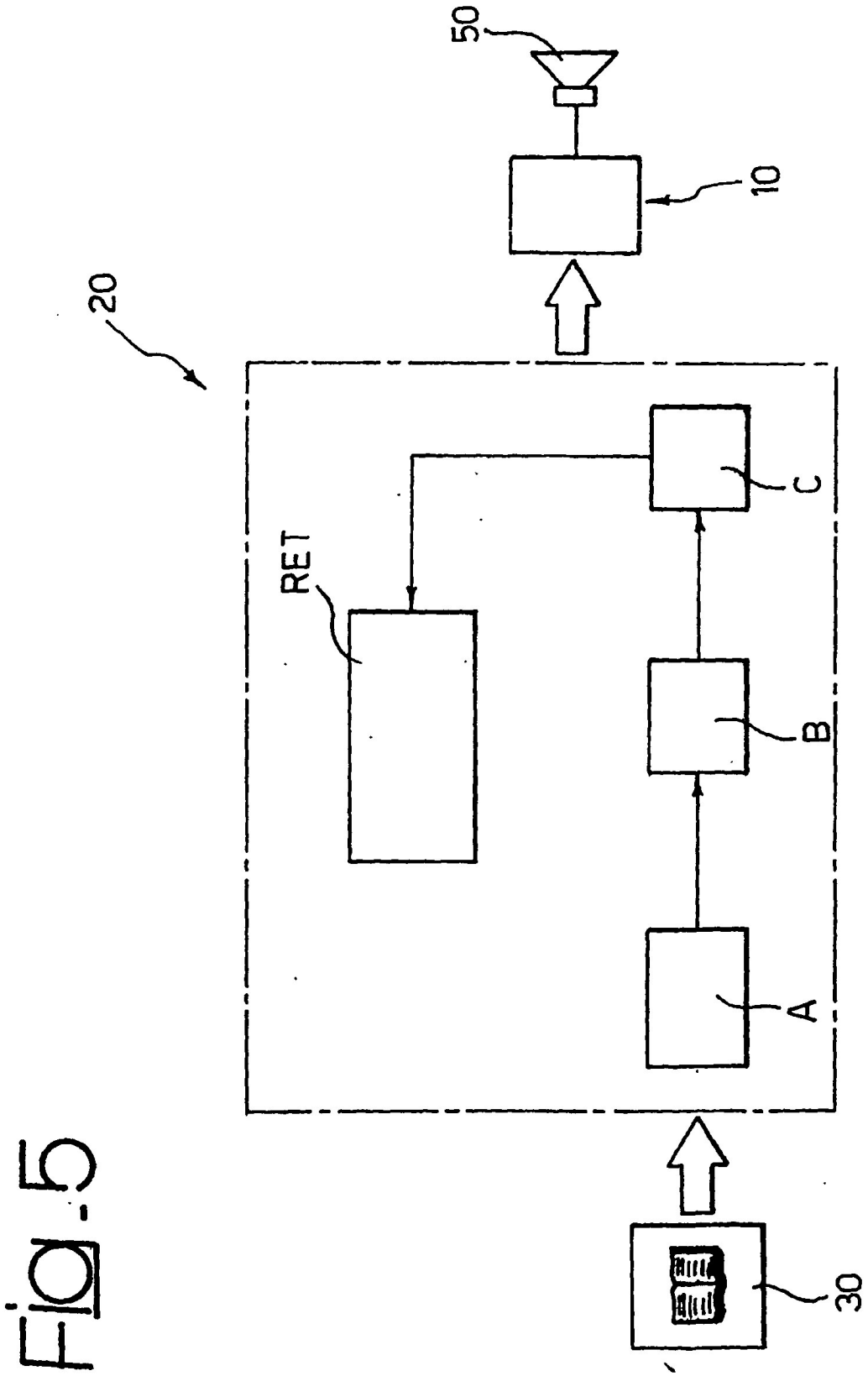


Fig. 5