(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 303 136**

(51) Int. Cl.:
G10L 15/06 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Número de solicitud europea: **04816199 .6**
(86) Fecha de presentación : **16.09.2004**
(87) Número de publicación de la solicitud: **1665231**
(87) Fecha de publicación de la solicitud: **07.06.2006**

(54) Título: **Procedimiento para el dopaje (introducción de impurezas) no supervisado y para el rechazo de palabras fuera de vocabulario en el reconocimiento vocal.**

(30) Prioridad: **16.09.2003 FR 03 50544**

(45) Fecha de publicación de la mención BOPI:
01.08.2008

(45) Fecha de la publicación del folleto de la patente:
01.08.2008

(73) Titular/es: **Telisma**
Parc d'Activité de Beaujardin
35410 Chateaugiron, FR

(72) Inventor/es: **Soufflet, Frédéric;**
Bleas, Gildas y
Cogne, Laurent

(74) Agente: **Elzaburu Márquez, Alberto**

DESCRIPCIÓN

Procedimiento para el dopaje (introducción de impurezas) no supervisado y para el rechazo de palabras fuera de vocabulario en el reconocimiento vocal.

La presente invención se refiere al ámbito de las interfaces vocales.

La presente invención se refiere de modo más particular a un motor de reconocimiento vocal, independientemente del contexto de la aplicación vocal particular: sistema de reconocimiento de la palabra para servidor telefónico, dictado vocal, sistema de mando y control embarcado, indexación de registros...

Los softwares comerciales actuales de reconocimiento de la palabra están basados en la utilización de redes ocultas de Markov (HMM, de Hidden Markov Model en inglés) para describir el vocabulario que hay que reconocer, y en una descodificación que utiliza un algoritmo de tipo Viterbi para asociar a cada enunciado una frase de este vocabulario.

Las redes markovianas en cuestión utilizan casi siempre estados de densidad continua, denominados frecuentemente GMM, de Gaussian Mixture Model.

El vocabulario de la aplicación, que éste basado inicialmente en gramáticas o en modelos de lenguaje estocásticos, es compilado en una red de estados finitos, con un fonema del idioma utilizado en cada transición de la red. El cambio de cada uno de estos fonemas por una red markoviana elemental, que representa este fonema en su contexto de coarticulación, produce finalmente una gran red markoviana a la cual se aplica la descodificación de Viterbi. Las redes elementales se han aprendido a su vez, gracias a un cuerpo de aprendizaje y con un algoritmo de aprendizaje bien conocido actualmente, por ejemplo de tipo Baum-Welsh.

La técnica anterior conoce ya por la patente europea EP 060 1778 “Clasificación de palabras-clave/de palabras no clave en el reconocimiento del lenguaje por palabras aisladas” un procedimiento para establecer si una señal de palabra que comprende un lenguaje digital representa o no una palabra-clave. La señal de palabra es tratada previamente en vectores característicos por un detector de palabras-clave por Modelo oculto de Markov (HMM). El detector de palabras-clave por Modelo oculto de Markov tiene como salidas señales que representan informaciones de segmentación de lenguaje y señales que representan notas de un conjunto de palabras-clave comparadas con el lenguaje. Las señales que representan los vectores característicos de la palabra hablada son tratadas a continuación utilizando un análisis por descenso probabilista generalizado y un procedimiento de combinación por discriminación lineal con miras a desarrollar un conjunto de notas de confianza, que, en comparación con un umbral, permite determinar si la palabra-clave es detectada.

De modo igualmente conocido, después de una fase de extracción acústica, la señal de palabra, que conceptualmente es una sucesión de fonemas continua o interrumpida por pausas, silencios o ruidos, es descrita por una sucesión de vectores acústicos. El trabajo de reconocimiento consiste en determinar qué sucesión de fonemas está asociada, de modo más probable, a esta sucesión de vectores acústicos, estando limitada esta búsqueda por el modelo de lenguaje de la aplicación.

El algoritmo de Viterbi es igualmente conocido por la técnica anterior. Se trata de determinar qué recorrido dentro de la red markoviana de la aplicación, es el recorrido más probablemente asociado a una sucesión de vectores acústicos, o bien decidir que el enunciado no es una de las frases para cuyo reconocimiento está construida la aplicación. Más bien que retener únicamente la sucesión de estados que ha obtenido el mejor resultado, es posible conservar varias de éstas, teniendo cuidado en tomar en consideración solamente las sucesiones que efectivamente están asociadas a enunciados diferentes (y no a variantes de un mismo enunciado, por ejemplo, con alineaciones temporales diferentes o bien con variantes de pronunciación diferentes). Esta técnica, tipo de descodificación Nbest, puede utilizarse para obtener los N mejores candidatos, cada uno, con un resultado asociado, siendo este resultado tanto más elevado cuanto más probable es la frase. El algoritmo precedente facilita una lista de candidatos con resultados asociados a cada uno, y falta determinar si realmente es cierto que el usuario ha pronunciado efectivamente uno de estos enunciados, o si se trata de otro fonéticamente parecido, o incluso de un ruido ambiente, véase, por ejemplo, el documento WO 0152239.

La experiencia muestra que la utilización del resultado obtenido por cada secuencia no es un buen índice para conservar una hipótesis, o bien rechazarla.

Una técnica particular es, por tanto, utilizar una red markoviana particular para absorber los enunciados fuera de vocabulario y los ruidos más frecuentes.

De modo preciso, este modelo de rechazo está constituido por una red markoviana que representa los diferentes ruidos posibles previstos por la aplicación, puesta en paralelo con una red markoviana constituida por el encadenamiento de cualquier sucesión de fonemas para el idioma utilizado, en número cualquiera.

Esta red está a su vez puesta en paralelo con la red efectiva de la aplicación.

Si la descodificación de Viterbi remonta un recorrido de esta red como mejor candidato, entonces el enunciado es rechazado.

Puede saberse, entonces, teóricamente, si este enunciado rechazado es ruido, o una palabra fuera de vocabulario no prevista por la aplicación.

Esta gestión del rechazo es eficaz cuando se trata de reconocer una palabra en una lista, pero es poco consistente para frases más complejas. Por ejemplo, si el usuario comienza con un principio de frase correcto (es decir, previsto por la gramática), y después utiliza palabras desconocidas para el sistema, entonces es probable que el mejor resultado no se obtenga con el modelo de rechazo, sino con una de las frases gramaticalmente previstas por el sistema, y que más se aproxime a la frase efectivamente pronunciada.

Por ejemplo, con un vocabulario basado en los días de la semana en francés (lundi, mardi,... dimanche), un locutor que pronuncie la palabra samedi (sábado) recalcando particularmente la segunda parte de la palabra (sameuuuudi), será reconocida como jeudi (jueves en francés). El resultado de la parte final de la palabra será optimizado por una asociación al final de la palabra jeudi, y a pesar de su comienzo completamente diferente, esta contribución al resultado final será suficiente para hacer bascular el algoritmo de Viterbi hacia un reconocimiento falso.

La técnica anterior conoce también técnicas de dopaje (introducción de impurezas) que tienen por objeto optimizar las tasas de reconocimiento, no de manera general por variantes a nivel de los métodos generales ya descritos, sino mejorar puntualmente para una aplicación dada, las prestaciones de reconocimiento. Cuando una aplicación de reconocimiento vocal es puesta en servicio, con frecuencia es posible recoger registros de las interacciones que han tenido lugar en el transcurso de los diálogos con el sistema.

Estando definido el modelo de lenguaje, es posible, escuchando estos registros, separar los que no tienen ninguna relación con la aplicación (diálogos fuera de vocabulario), y clasificar los otros en una de las categorías siguientes:

- Frases correctamente reconocidas
- Frases completa o parcialmente falsas (error de reconocimiento).

La idea del dopaje es utilizar estos registros para mejorar los modelos acústicos generales del idioma para una aplicación dada, teniéndolos en cuenta en un nuevo aprendizaje, que está basado en un principio parecido al aprendizaje de Baum-Welsh que ha permitido obtener los modelos iniciales generalistas. Pero esta vez, los registros corresponden exactamente a las palabras utilizadas por la aplicación, en condiciones similares a las de su utilización. Se obtienen entonces reducciones importantes de las tasas de error.

El dopaje supervisado es un método manual de puesta en práctica del principio general del dopaje que acaba de presentarse.

Hay que escuchar los registros y clasificarlos dentro de uno de los conjuntos mencionados anteriormente.

En las frases completa o parcialmente falsas, hay que determinar entonces si se trata de una sustitución (la frase realmente pronunciada, no es la reconocida, sino que forma parte de las frases admitidas por el modelo de lenguaje de la aplicación). Si éste es el caso, se modifica el texto asociado. Si no es el caso, se la descarta (aunque en ciertas variantes, a pesar de todo, ésta pueda ser utilizada). Esta frase podrá servir, por ejemplo, para mejorar el rechazo.

Así pues, el dopaje supervisado es una operación manual, por tanto, larga y costosa. Típicamente, para mejorar de modo visible la aplicación, se necesitan varios millares de registros. Por el contrario, éste es un método completamente eficaz, y es típico reducir en ciertos casos las tasas de error en un factor de 2.

Como el dopaje supervisado es muy costoso, es posible utilizar un dopaje no supervisado.

En este caso, se comienza por regular de modo severo y, por tanto, rechazar un enunciado dudoso más bien que arriesgarse a un error de sustitución.

A continuación se utilizan los registros como anteriormente, sin verificar por una escucha efectiva que la frase reconocida es la pronunciada.

Este método poco costoso mejora típicamente la tasa de reconocimiento en frases que estaban ya correctamente reconocidas, pero refuerza las tasas de error en las frases que planteaban ya problema.

Este es la razón por la cual esta técnica es poco utilizada, o es mezclada entonces con ciertas operaciones supervisadas.

La presente invención pretende remediar los inconvenientes de la técnica anterior, proponiendo un procedimiento nuevo que permita utilizar con éxito el dopaje no supervisado y que, igualmente, aumente de modo muy sensible la eficacia del tratamiento del rechazo en reconocimiento vocal.

Para hacer esto, la presente invención, tal como se reivindica en la reivindicación 1, es del tipo descrito anteriormente, y destaca, en su acepción más amplia, porque se basa en un procedimiento de reconocimiento de palabra que

comprende una etapa de descodificación de tipo Viterbi en una red de Markov, para obtener una lista de N candidatos, caracterizado porque se determina para cada fonema fuente Ps un conjunto asociado EAPD de fonemas Pd muy diferenciables con el citado fonema fuente Ps, y porque cada uno de los citados candidatos es objeto de un tratamiento consistente en comparar cada uno de estos fonemas con los fonemas del citado conjunto asociado EAPD, y en rechazar el citado candidato si, al menos, uno de los fonemas evaluados es confundido con uno de los citados fonemas del citado conjunto asociado.

Ventajosamente, la presente invención comprende una etapa en la que los candidatos no rechazados son objeto de un tratamiento adicional, consistente en determinar para cada fonema fuente Ps un conjunto asociado EAPF de fonemas Pf poco diferenciables con el citado fonema fuente Ps, y en la que cada uno de los citados candidatos no rechazados durante las etapas precedentes es objeto de un tratamiento consistente en comparar cada uno de sus fonemas con los fonemas del citado conjunto asociado EAPF y en excluir de las etapas de dopaje las porciones de la señal correspondientes al fonema confundido, sin rechazar al citado candidato.

Ventajosamente, la presente invención comprende una etapa adicional de tratamiento de los candidatos no rechazados, consistente en determinar para cada fonema fuente Ps un conjunto asociado EAPM de fonemas Pm medianamente diferenciables con el citado fonema fuente Ps, y en la que cada uno de los citados candidatos no rechazados es objeto de un tratamiento consistente en comparar cada uno de sus fonemas con los fonemas del citado conjunto asociado EAPM y en rechazar el citado candidato si el fonema es confundido y si, al menos, una de las confusiones evaluadas corresponde a una palabra válida del diccionario.

Preferentemente, las porciones de señal no excluidas son utilizadas para un reaprendizaje del citado modelo Markoviano y para una reestimación de los conjuntos EAPS, EAPM y EAPF, y de los comparadores.

Preferentemente, los tratamientos son repetidos.

La invención se comprenderá mejor con la ayuda de la descripción que seguidamente se hace a título puramente explicativo, de un modo de realización de la invención, refiriéndose a las figuras anejas:

La figura 1 ilustra un esquema que describe los cuatro módulos del procedimiento de acuerdo con la invención.

La figura 2 ilustra una red elemental markoviana de la [a] en francés de acuerdo con el estado de la técnica.

La figura 3 ilustra una red elemental markoviana modificada de la [a] en francés de acuerdo con la invención.

La figura 4 ilustra un ejemplo de matriz de diferenciación por Multi Layer Perceptrons (Perceptrones Multicapa) (MLP) de acuerdo con la invención.

De acuerdo con la invención, ilustrada en la figura 1, el primer módulo utiliza una descodificación clásica de tipo Nbest markoviana para la obtención de las N mejores asociaciones posibles entre el enunciado del usuario y un conjunto de hipótesis plausibles que provienen del modelo de lenguaje de la aplicación.

Las redes markovianas elementales clásicas del idioma utilizado son modificadas para dar a los destinos de las redes propiedades ventajosas que se van a describir. Supóngase, por ejemplo, que la red elemental markoviana de la [a] en francés, de acuerdo con el estado de la técnica, tiene una topología ilustrada en la figura 2. Se va a modificar entonces esta estructura y a utilizar una red diferente, para aislar un número reducido de densidades.

El objeto es que, en el transcurso de una descodificación, e independientemente de los contextos acústicos izquierdos y derechos, se obligue a la alineación final a pasar, al menos, una vez por, al menos, uno de los destinos definidos.

Este objeto puede obtenerse, por ejemplo, gracias a la topología ilustrada en la figura 3.

Se ve que con esta topología, se pasa obligatoriamente por una de las tres densidades: fh2, fn2 o fv2. La primera corresponde a la aspiración, la segunda a la nasalización, y la tercera a la vocalización. Esta elección es ilustrativa, lo importante es aislar un pequeño número de densidades destino, que formen un grupo en el cual, al menos, una sea atravesada forzosamente por una alineación que utiliza el fonema, cualquiera que sea el contexto. Modificaciones basadas en el mismo principio deben ser aportadas a todos los fonemas del idioma.

Como resultado, cada una de las alineaciones obtenidas a la salida del primer módulo, contiene, para cada fonema de la secuencia, al menos una ocurrencia de un representante de los grupos de los destinos que se han definido.

El segundo módulo es un módulo nuevo de validación o de invalidación de estas hipótesis con la ayuda de métodos no markovianos. En una implementación particular, éste está basado en los Multi Layers Perceptrons (MLP). Los MLPs están definidos para cada par de fonemas del idioma y son activados para separar lo mejor posible estos fonemas. Estos son comparadores. Estos MLPs se utilizan para eliminar las hipótesis de la lista de los 15-20 mejores, presentando, por tanto, un resultado elevado, pero que no satisface condiciones fonológicas más estrictas de aceptación.

Este módulo, puede, así, decidir rechazar una o varias frases hipótesis, de modo directo, y sin hacer referencia al resultado obtenido por esta hipótesis en el primer módulo markoviano.

De modo más preciso, el primer módulo de descodificación markoviano produce, para cada una de las soluciones concurrentes, una alineación que pone en relación las tramas acústicas con estados markovianos de los modelos elementales asociados a los fonemas.

El interés está, entonces, en las tramas que están asociadas a ciertos fonemas que tienen propiedades particulares, principalmente de ser raramente confundidos con otros fonemas. Los MLPs son activados entonces en estos fonemas.

Un ejemplo típico es el grupo de vocales, que son estables y energéticas, por tanto poco sensibles a los ruidos y, al menos, en algunas, fácilmente separables entre ellos.

Para cada par de fonemas del idioma, puede definirse, entonces, un porcentaje de confusión del MLP correspondiente. Se obtiene, así, una matriz simétrica denominada matriz de diferenciación, como está representado en la figura 4.

Por ejemplo, si se toma la columna correspondiente al fonema [o], se ve que el MLP asociado, en el cuerpo de palabra utilizado para producir esta tabla, no ha confundido nunca una [o] con una [y] (caso 0), pero ha confundido una [o] con un [un] 3,80 veces de cien (caso 3,80).

Para un fonema fuente Ps fijado, se aíslan entre todos los MLPs de la tabla, en un primer conjunto EAPD de fonemas Pd, aquéllos que correspondan a tasas de confusión entre fonemas inferiores a un umbral bajo, por ejemplo el 3%. Estos fonemas Pd se denominan muy diferenciables con el fonema fuente Ps.

Se conservará, por ejemplo, el MLP correspondiente al par [o][y], pero no el correspondiente al par [o][un].

Este conjunto de MLPs se utiliza, entonces, del modo siguiente:

> Para cada hipótesis producida por el primer módulo y su alineación, se verifica que todos los MLPs del conjunto descrito anteriormente, en el cual este fonema aparece, clasifican bien los vectores acústicos asociados por alineación en la categoría adecuada.
>
> Si uno de los MLPs del conjunto produce un error de clasificación en uno de los fonemas de la alineación, entonces la hipótesis es rechazada.

Si todas las hipótesis que proceden del primer módulo son rechazadas, entonces la frase pronunciada será clasificada fuera de vocabulario y rechazada.

Las hipótesis que han resistido a este segundo módulo son candidatos para el dopaje no supervisado, y son transmitidas al tercer módulo.

En un ejemplo particular de implementación, para cada par de grupo de destinos, se va a definir un MLP, que será activado para separar lo mejor posible dos fonemas.

En este ejemplo, una red está constituida por una capa de entrada de dimensión 120, una capa oculta de dimensión 15 y una capa de salida de dimensión 2, donde todas las células de una capa están unidas a todas las células de la capa precedente.

En un ejemplo particular de implementación, el vector de entrada, de dimensión 120, se forma del modo siguiente:

> Se integran las energías en un banco de 24 filtros triangulares repartidos regularmente en la escala Mel entre 188 Hz y 3594 Hz, en la ventana acústica asociada al grupo de destinos, para la alineación utilizada. La escala Mel retenida es la siguiente:

$$f_{Mel} = 2595.\log_{10}(1+\frac{f_{Hz}}{700})$$

Se obtiene, así, un vector de 24 coeficientes V3. Se normaliza el vector V3. Se efectúa la misma operación utilizando la ventana que precede para obtener V2, después la que precede aún más para obtener V1. Igualmente, se calcula V4 por el mismo procedimiento en la ventana siguiente, después V5 en la ventana que sigue. Se compone, entonces, el vector de entrada por concatenación de [V1 V2 V3 V4 V5].

En una variante de implementación, se podrán normalizar los vectores V1, V2, V4 y V5 con el coeficiente de normalización obtenido para el vector V3.

En otra variante de implementación, si la alineación muestra que quedan x transiciones del grupo de destinos, el vector V3 podrá calcularse a partir de estos x vectores acústicos (podrá utilizarse, por ejemplo, la media de las energías de cada una de las bandas). Los vectores V2 y V1 y V4 y V5 se calcularán utilizando, respectivamente, las ventanas antes de la primera transición del grupo de destinos y las ventanas después de la última transición del grupo de destinos.

Las células de entrada forman solamente una copia del vector de entrada.

Para cada célula de la capa oculta y de la capa de salida, existe un umbral.

Se puede hacer aprender los pesos de las relaciones y los umbrales de las células de la red por presentación de un cuerpo de vectores de entrada y de las salidas deseadas correspondientes, y retropropagación de gradiente (descenso de gradiente de capa en capa).

Por aprendizaje, se busca un mínimo del error medio cuadrático entre las transformadas de los vectores de aprendizaje por la red y los vectores destino correspondientes. Así pues, se obtienen nubes de puntos centradas alrededor de cada uno de los 2 vectores destino, correspondiendo cada uno a una clase. La discriminación es significativa si los ensanchamientos de estas nubes son pequeños frente a las distancias entre los vectores destino. La red óptima es aquélla que da el error medio cuadrático más pequeño en la base de prueba.

El tercer módulo recibe las hipótesis que no han sido eliminadas por el segundo módulo.

Las hipótesis que se van a tratar en este módulo, formaban parte de la lista Nbest del primer módulo y, por otra parte, la alineación que está asociada a estas, es coherente en el sentido de que ninguno de los fonemas que la componen es "groseramente" invalidado en el segundo módulo por los MLPs especializados en la detección de las confusiones más improbables desde un punto de vista acústico.

Así pues, estas hipótesis son trivialmente aberrantes. Si se considera que estas hipótesis no corresponden a la frase realmente pronunciada, se trataría con toda probabilidad de un caso de sustitución de una palabra fuera del vocabulario existente en el idioma, en lugar de la palabra prevista por la gramática de la aplicación. La parte común a las dos frases podrá utilizarse, entonces, sin problema para el dopaje de la aplicación, incluso si la misma frase deberá ser rechazada.

El objeto del tercer módulo es determinar si se está efectivamente en un caso de sustitución de este tipo.

Si éste no es el caso, se toma como resultado del reconocimiento la hipótesis de mayor resultado restante después del tratamiento del segundo módulo, y las otras eventuales hipótesis de menor resultado serán las alternativas posibles, menos probables.

Si, por el contrario, existe una presunción de sustitución local, entonces la función de este tercer módulo será determinar qué partes de la frase pronunciada y, de modo más preciso, qué partes de la secuencia de vectores acústicos calculados para esta frase, deberán ser retenidas, y para esta parte, cuáles son los fonemas que hay que considerar para el propio dopaje.

Con este objeto, se define un segundo conjunto de MLPs que esta vez contiene los MLPs que presentan las mayores tasas de confusión posibles entre dos fonemas.

Así, para un fonema fuente Ps fijado, se aíslan entre todos los MLPs de la tabla, en un conjunto EAPF de fonemas Pf aquéllos que corresponden a tasas de confusión entre fonemas superiores a un umbral alto, por ejemplo el 10%. Estos fonemas Pf se denominan poco diferenciables con el fonema fuente Ps.

Con la matriz de la figura 4, se tendrán, entonces, por ejemplo, los pares siguientes:

[z][s]: 12,08%

[t][k]: 16,15%

[in][un]: 19,13%

[s][f]: 22,44%

Si una de las alineaciones conduce a una sola confusión entre fonemas de estos pares, entonces la hipótesis se considerará como válida y no rechazada.

Por el contrario, los vectores acústicos correspondientes a las tramas erróneas en esta sustitución no serán retenidos en la alineación.

Supóngase que la palabra pronunciada, cuando es tratada por la primera etapa markoviana, vuelve a la sucesión Nbest siguiente (con N = 3, por ejemplo).

Cero: resultado 1200

Seis: resultado 200

Ocho: resultado 50

Supóngase que el segundo módulo, utilizando, por ejemplo, los MLPs basados en los pares que contienen la [i], elimina las dos últimas hipótesis, pero la primera es validada.

Supóngase, entonces, que en el tercer módulo, el MLP asociado al par [s][z] permite concluir que el primer fonema es más probablemente una [s] que una [z].

Como este MLP forma parte del grupo especial de altas tasas de confusión que se han descrito, la hipótesis será conservada, pero las tramas asociadas a la [z] serán eliminadas de la secuencia para el dopaje.

Así, el registro aportará una contribución al dopaje de los tres fonemas.

[e][r][o], siendo tomado el contexto de la [e] entre [z] y [e].

Finalmente, el tercer módulo utiliza, también, un tercer conjunto EAPM de fonemas Pm para cada fonema fuente Ps, a saber, aquéllos que no forman parte de los dos primeros conjuntos EAPD y EAPF y, que, por tanto, representan pares de fonemas con probabilidades de confusión medias. Esos fonemas son denominados medianamente diferenciables.

Por otra parte, en una variante de implementación de la invención, puede seleccionarse en este tercer conjunto un solo subconjunto de los MLPs restantes. Para el funcionamiento del descodificador de acuerdo con el principio presentado aquí, no es necesario que todos los MLPs sean utilizados al final.

Además de este tercer conjunto de MLPs, el tercer módulo utiliza un diccionario fonético de las palabras del idioma, con sus características sintácticas principales (nombre masculino o femenino, por ejemplo en francés).

Este diccionario contiene, igualmente, un índice de frecuencia de la palabra en la utilización corriente de este idioma. Por ejemplo, la palabra “maison” es más frecuente que la palabra “maistrance”.

Cuando uno de los MLPs del conjunto detecta una confusión posible entre dos fonemas de un par, este MLP producirá, también, un resultado de probabilidad para esta confusión (este resultado de probabilidad de confusión será función del vector de salida del MLP y del ensanchamiento medio alrededor de los valores destino durante la fase de aprendizaje).

Este resultado será multiplicado por el índice de frecuencia de la palabra, siendo comparado el resultado con un umbral fijado. 10 es un valor estándar de este umbral cuando la tasa de frecuencia de una palabra es un coeficiente comprendido entre 0 (palabra muy rara en el contexto de la aplicación) y 100 (palabra muy frecuente en el ámbito de la aplicación, pero, sin embargo, fuera de vocabulario).

La palabra será rechazada si una de las palabras del diccionario presenta una sustitución posible con un resultado superior al umbral.

Por el contrario, una vez más, los fonemas que no están asociados al MLP poniendo en evidencia una sustitución posible serán conservados para el dopaje.

El cuarto módulo es un módulo de dopaje.

Cuando suficientes tramas hayan sido asociadas a un fonema o a un grupo predefinido de fonemas, y no hayan sido rechazadas por ninguno de los módulos precedentes, es posible utilizar esta asociación trama fonema para reevaluar los valores de las densidades gausianas, por uno de los métodos estándar descritos.

Asimismo, es posible reevaluar cada uno de los coeficientes de los MLPs utilizados.

Así, con intervalos regulares, se actualizarán las redes de Markov del primer módulo y los MLPs de los módulos siguientes, permitiendo, de esta manera, obtener un sistema cada vez más eficaz en tasa de reconocimiento.

La invención se ha descrito en lo que precede a título de ejemplo. Naturalmente, el experto en la técnica es capaz de realizar diferentes variantes de la invención sin por ello salirse del marco de la patente.

REIVINDICACIONES

1. Procedimiento de reconocimiento de palabra que comprende una etapa de descodificación de tipo Viterbi en una red de Markov, para obtener una lista de N candidatos, **caracterizado** porque se determina para cada fonema fuente Ps un conjunto asociado EAPD de fonemas Pd muy diferenciables con el citado fonema fuente Ps, y porque cada uno de los citados candidatos es objeto de un tratamiento consistente en comparar cada uno de estos fonemas con los fonemas del citado conjunto asociado EAPD, y en rechazar el citado candidato si, al menos, uno de los fonemas evaluados es confundido con uno de los citados fonemas del citado conjunto asociado.

2. Procedimiento de reconocimiento de acuerdo con la reivindicación 1, **caracterizado** porque los candidatos no rechazados son objeto de un tratamiento adicional, consistente en determinar para cada fonema fuente Ps un conjunto asociado EAPF de fonemas Pf poco diferenciables con el citado fonema fuente Ps, y porque cada uno de los citados candidatos no rechazados durante las etapas precedentes es objeto de un tratamiento consistente en comparar cada uno de sus fonemas con los fonemas del citado conjunto asociado EAPF y en excluir de las etapas de dopaje las porciones de la señal correspondientes al fonema confundido, sin rechazar al citado candidato.

3. Procedimiento de reconocimiento de acuerdo con las reivindicaciones 1 o 2, **caracterizado** porque comprende una etapa adicional de tratamiento de los candidatos no rechazados, consistente en determinar para cada fonema fuente Ps un conjunto asociado EAPM de fonemas Pm medianamente diferenciables con el citado fonema fuente Ps, y porque cada uno de los citados candidatos no rechazados es objeto de un tratamiento consistente en comparar cada uno de sus fonemas con los fonemas del citado conjunto asociado EAPM y en rechazar el citado candidato si el fonema es confundido y si, al menos, una de las confusiones evaluadas corresponde a una palabra válida del diccionario.

4. Procedimiento de reconocimiento de acuerdo con la reivindicación 3, **caracterizado** porque las porciones de señal no excluidas se utilizan para un reaprendizaje del citado modelo Markoviano y para una reestimación de los conjuntos EAPS, EAPM y EAPF, y de comparadores.

5. Procedimiento de reconocimiento de acuerdo con las reivindicaciones precedentes, **caracterizado** porque los tratamientos son repetidos.

6. Procedimiento de reconocimiento de acuerdo con las reivindicaciones 1, 2 o 3, **caracterizado** porque los citados conjuntos EAPD, ESPM y EAPF se obtienen a partir de una matriz de diferenciación.

7. Procedimiento de reconocimiento de acuerdo con la reivindicación 6, **caracterizado** porque los coeficientes de la citada matriz de diferenciación son las tasas de errores de comparadores Multi Layers Perceptrons asociados a un par de fonemas.

8. Procedimiento de reconocimiento de acuerdo con una de las reivindicaciones 6 o 7, **caracterizado** porque el citado conjunto EAPF es, para un fonema fuente Ps dado, el conjunto de los fonemas cuyas tasas de diferenciación de la citada matriz de diferenciación con el citado fonema fuente son superiores a un umbral alto.

9. Procedimiento de reconocimiento de acuerdo con una de las reivindicaciones precedentes, **caracterizado** porque el citado conjunto EAPD es, para un fonema fuente Ps dado, el conjunto de los fonemas cuyas tasas de diferenciación de la citada matriz de diferenciación con el citado fonema fuente son inferiores a un umbral bajo.

10. Procedimiento de reconocimiento de acuerdo con una de las reivindicaciones precedentes, **caracterizado** porque el citado conjunto EAPM es, para un fonema fuente Ps dado, el conjunto de los fonemas cuyas tasas de diferenciación de la citada matriz de diferenciación con el citado fonema fuente están comprendidas entre el citado umbral alto y el citado umbral bajo.

11. Procedimiento de reconocimiento de acuerdo con las reivindicaciones 9 y 10, **caracterizado** porque el citado umbral bajo de diferenciabilidad es de aproximadamente el 3% y el citado umbral alto de diferenciabilidad es de aproximadamente el 10%.

12. Procedimiento de reconocimiento de acuerdo con la reivindicación 3, **caracterizado** porque el citado diccionario es un diccionario fonético de las palabras del idioma con sus características principales y un índice de frecuencia de la palabra en la utilización corriente del idioma.

13. Procedimiento de reconocimiento de acuerdo con la reivindicación 1, **caracterizado** porque la citada red de Markov es una red modificada para aislar un número reducido de densidades.

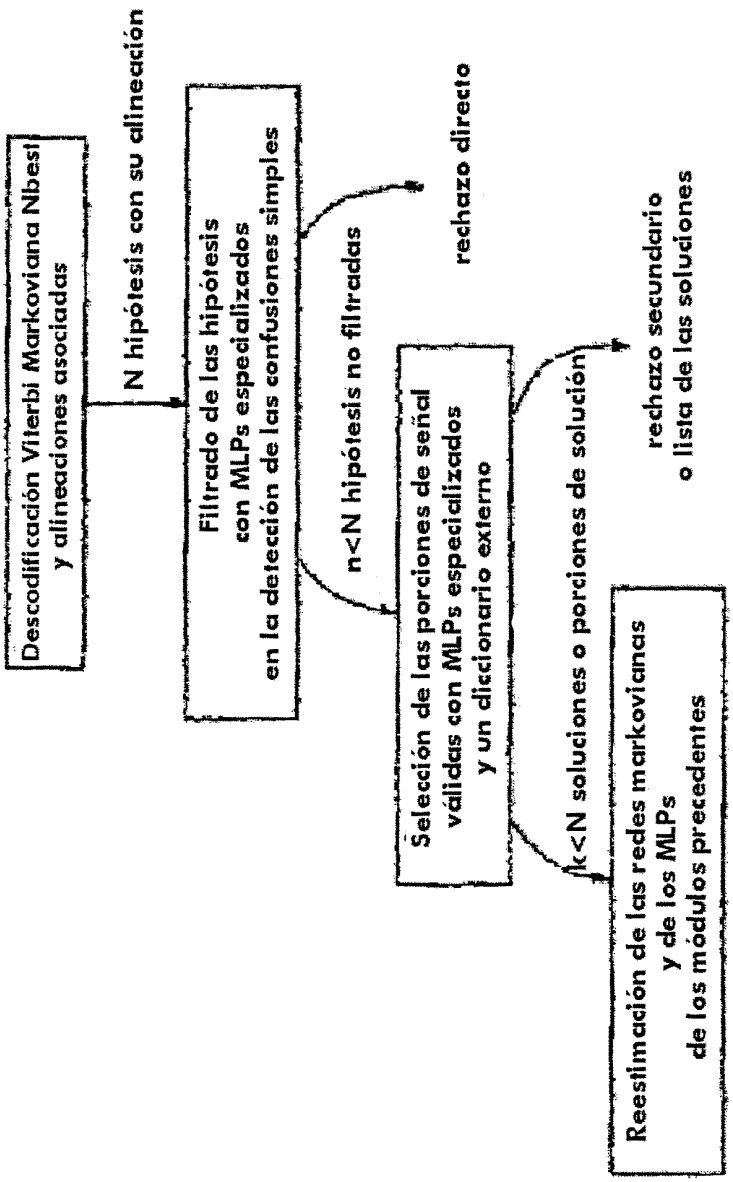
Descodificación Viterbi Markoviana Nbest
y alineaciones asociadas
N hipótesis con su alineación
Filtrado de las hipótesis
con MLPs especializados
en la detección de las confusiones simples
n<N hipótesis no filtradas
rechazo directo
Selección de las porciones de señal
válidas con MLPs especializados
y un diccionario externo
k<N soluciones o porciones de solución
rechazo secundario
o lista de las soluciones
Reestimación de las redes markovianas
y de los MLPs
de los módulos precedentes

Figura 1

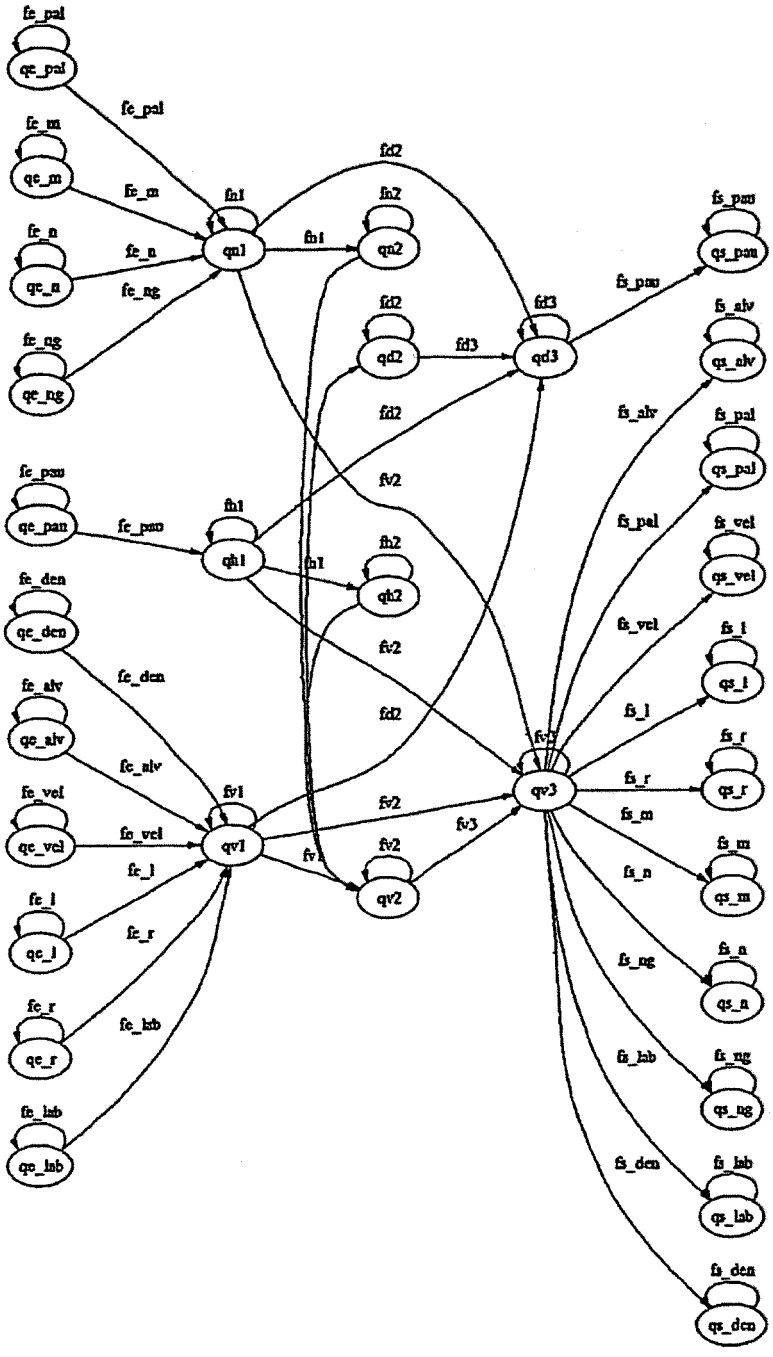
fe_pal
qe_pal
fe_pal
fe_m
qe_m
fe_m
fe_n
qe_n
fe_n
fe_ng
qe_ng
fe_ng
qn1
fn1
fn1
qn2
fn2
fd2
qd2
fd2
fd3
qd3
fd3
fd2
fv2
fe_pan
qe_pan
fe_pan
qh1
fh1
fh1
qh2
fh2
fv2
fe_den
qe_den
fe_den
fe_alv
qe_alv
fe_alv
fe_vel
qe_vel
fe_vel
fe_l
qe_l
fe_l
fe_r
qe_r
fe_r
fe_lab
qe_lab
fe_lab
qv1
fv1
fv1
fv2
qv2
fv2
fv2
fv3
qv3
fv3
fs_pan
qs_pan
fs_pan
fs_alv
qs_alv
fs_alv
fs_pal
qs_pal
fs_pal
fs_vel
qs_vel
fs_vel
fs_l
qs_l
fs_l
fs_r
qs_r
fs_r
fs_m
qs_m
fs_m
fs_n
qs_n
fs_n
fs_ng
qs_ng
fs_ng
fs_lab
qs_lab
fs_lab
fs_den
qs_den
fs_den

Figura 2

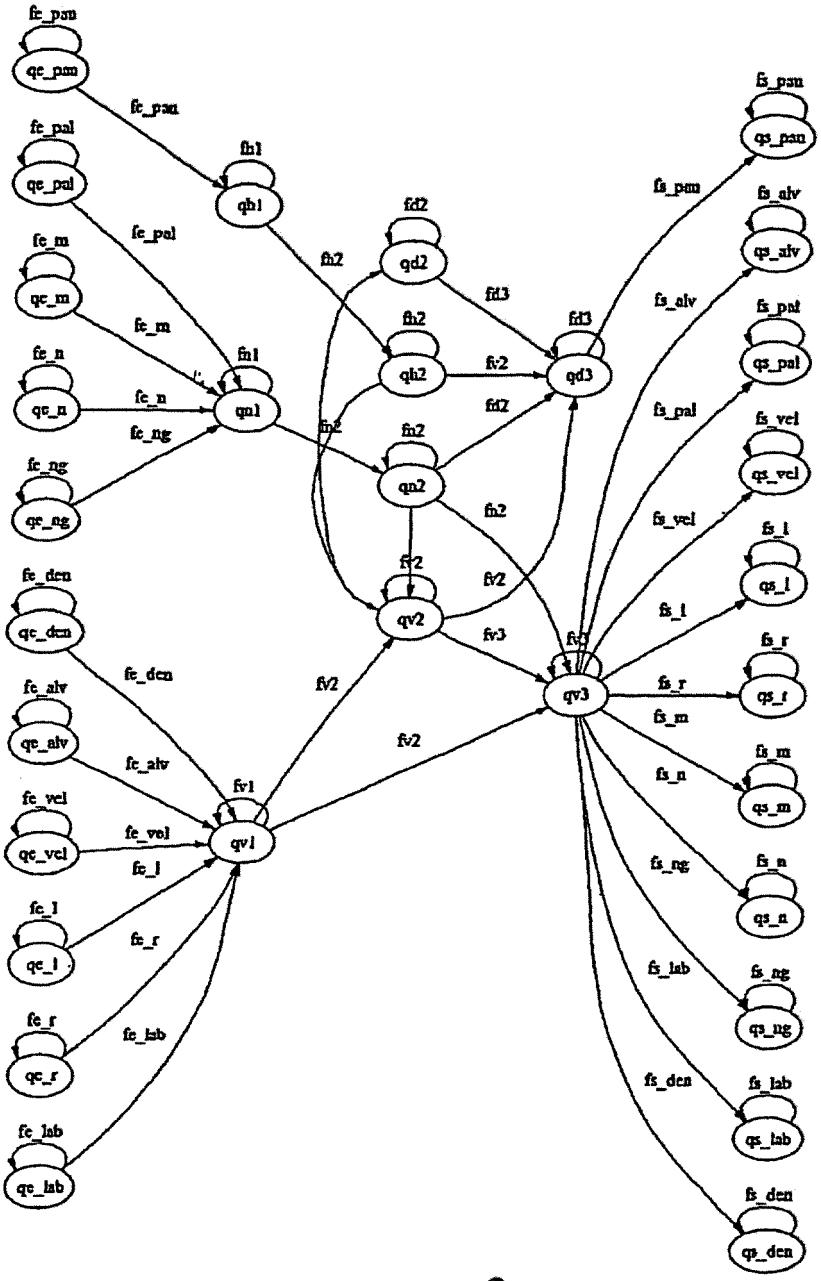
fc_pan
qc_pan
fc_pal
qc_pal
fc_m
qc_m
fc_n
qc_n
fc_ng
qc_ng
fc_den
qc_den
fc_alv
qc_alv
fc_vel
qc_vel
fc_l
qc_l
fc_r
qc_r
fc_lab
qc_lab
fh1
qh1
fn1
qn1
fv1
qv1
fh2
qd2
qh2
qn2
qv2
fd2
fd3
qd3
fv2
fv3
qv3
fs_pan
qs_pan
fs_alv
qs_alv
fs_pal
qs_pal
fs_vel
qs_vel
fs_l
qs_l
fs_r
qs_r
fs_m
qs_m
fs_n
qs_n
fs_ng
qs_ng
fs_lab
qs_lab
fs_den
qs_den

Figura 3

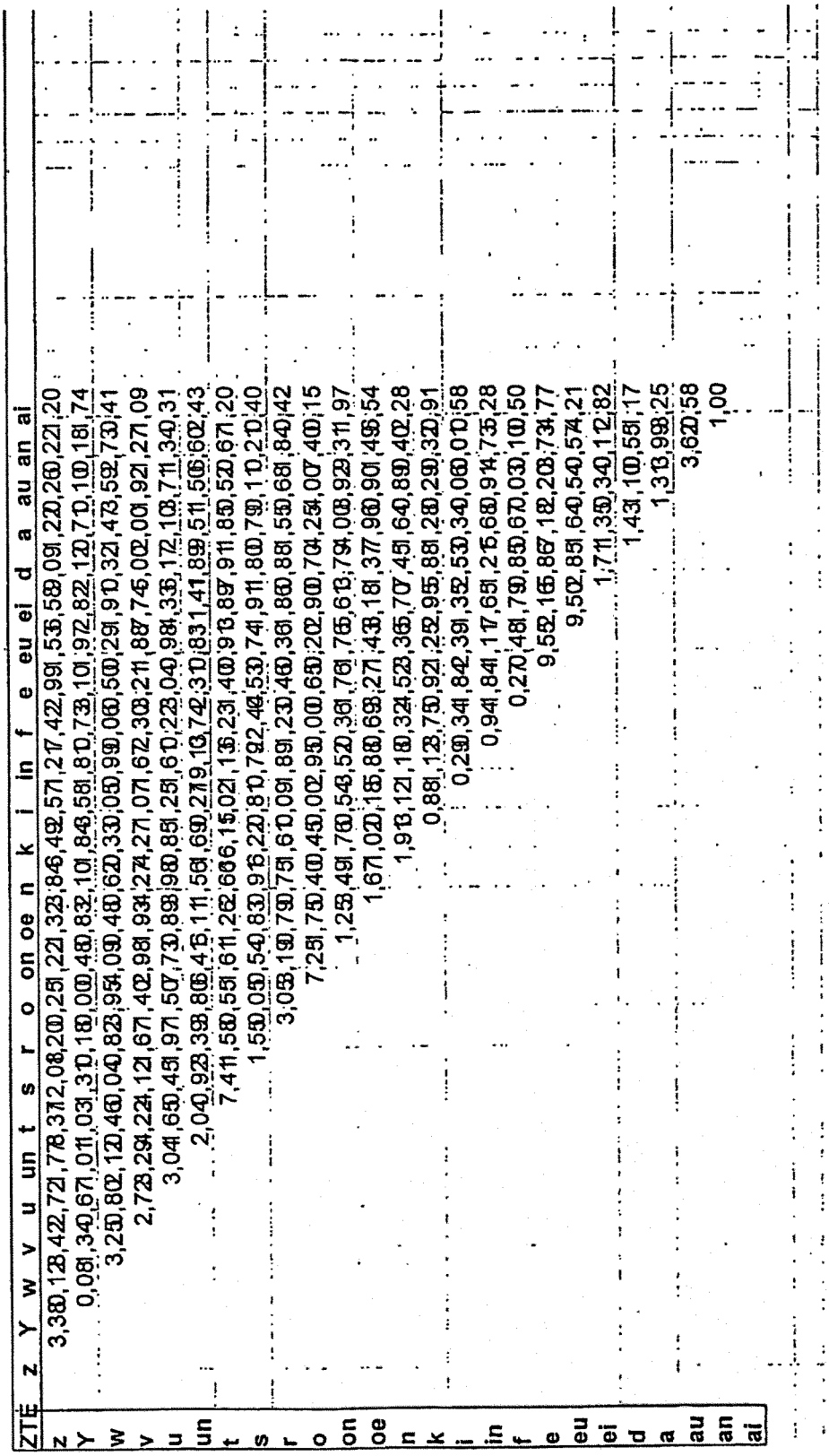

Figura 4