Universidad Nacional de Colombia, Seminario de Investigación
Abril de 2005

Estado del arte en el reconocimiento Automático de voz

Deiby Alexander Fandiño Rodríguez

Universidad Nacional de Colombia

Resumen

El Reconocimiento Automático de Voz (ASR Automatic Speech Recognition) es un campo de investigación de creciente relevancia que día a día se gana más adeptos. El desarrollo de mejores algoritmos y de modelados mas precisos, junto con la aparición de sistemas informáticos mas potentes y adsequibles, posibilita la integración de los sistemas de dialogo hombre-máquina a través de la voz en numerosos ámbitos de la sociedad actual. Estos sistemas de dialogo permiten el acceso a una gran cantidad de información a través de una forma de comunicación tan natural como es el habla, facilitando un elevado numero de servicios interactivos utilizando el teléfono, la televisión o el ordenador como elementos de acceso.

El objetivo general de este artículo es presentar los principales avances obtenidos en los últimos años en el ámbito del reconocimiento automático de voz. Se presta especial atención a las principales líneas de trabajo orientadas hacia el diseño de sistemas de Reconocimiento del Habla.

I. INTRODUCCIÓN

Hace ya tiempo que se estudia la posibilidad de desarrollar interfaces hombre-máquina controlados por la voz para sustituir en ciertas aplicaciones a los interfaces tradicionales basados en teclados, paneles y dispositivos similares. Este nuevo tipo de interfaz constaría de dos módulos de entrada/salida: uno de reconocimiento de habla, mediante el cual el ordenador sería capaz de extraer información de los comandos orales del operador o usuario, y otro de síntesis de voz, que podría ser una de las vías de presentación de resultados.

La utilización de la voz, y en el caso que nos ocupa, el Reconocimiento de Habla, como vía de dar órdenes a los ordenadores ofrece varias ventajas respecto al método tradicional de comunicación entre el usuario y la máquina:

Hace esta comunicación más rápida, y más agradable para los nuevos usuarios, ya que al ser la forma natural de comunicarse no se necesita ninguna habilidad especial.

Permite el tener las manos libres para utilizarlas en alguna otra actividad, a la vez que se van dando órdenes por medio de la voz.

Permite movilidad, ya que la voz se puede enviar a distancia y ser recogida por un micrófono, por oposición a un teclado que no se puede mover de la mesa de trabajo.

Permite acceso remoto, al poder acceder a un ordenador usando la red telefónica, que es la red de comunicaciones más extendida.

Permite la disminución del tamaño de los paneles de control. Piénsese en el panel de un avión, cuantos conmutadores manuales podrían suprimirse si se utilizara la voz como forma de comunicación con el sistema de control

A lo largo del presente artículo se pretende presentar una panorámica del problema del Reconocimiento del Habla, así como de las soluciones técnicas que hasta ahora se han desarrollado, acabando con una revisión de las posibles aplicaciones que pueda servir para despertar en aquellas personas que no estén al corriente del desarrollo de esta tecnología, interés por la misma así como vislumbrar posibles aplicaciones en sus propios campos de actividad.

II. JUSTIFICACIÓN

Durante la declinación del ya finalizado siglo XX y, por supuesto, continuando tras el nacimiento del tan esperado siglo XXI, la utilización cada vez mayor de la voz como interfaz de comunicación entre los hombres y las máquinas permite aumentar la cooperación con los sistemas informáticos, aprovechando al máximo las prestaciones de estos en cuanto a rapidez y eficiencia.

Los avances que se producen en el ámbito de las tecnologías del habla son día a día más significativos. En el campo del reconocimiento automático de voz, los reconocedores actuales manejan cada vez vocabularios más grandes y logran menores tasas de error gracias al uso de algoritmos más eficientes, a la aparición de equipos más

potentes y baratos, y al aumento de complejidad de estos sistemas, al emplearse modelados mas sofisticados y refinados.

Los sistemas de reconocimiento automático de voz o habla, frente a otros sistemas de interacción hombre-máquina como teclados, paneles, etc., proporcionan una mayor naturalidad, así como un amplio rango de utilización por parte de diferentes tipos de usuarios en distintos entornos de operación.

No obstante, a pesar de los grandes avances realizados, se está todavía muy lejos de un sistema de reconocimiento automático de voz universal que funcione bien en cualquier aplicación a la que sea destinado. En general, el diseño y las características de los actuales sistemas de reconocimiento automático de voz dependen fuertemente de la aplicación a la que van a ser destinados y a las condiciones de funcionamiento.

III. DEFINICIÓN DEL PROBLEMA

El Reconocimiento del Habla parece tan natural y sencillo para las personas que se pensó que podría ser fácilmente realizado por las máquinas. Sin embargo, cuando se empezó a profundizar en el tema, se comprobó que esto no es así. De hecho, es un tema que se ha revelado más complicado que la producción automática de voz.

Ya la historia lo ha demostrado: las primeras y rudimentarias máquinas parlantes aparecieron en la segunda mitad del siglo XVIII, mientras que los primeros intentos en máquinas capaces de reconocer la voz no aparecieron hasta principios del siglo XX, con la máquina de Flower, capaz de escribir el alfabeto fonográfico pronunciado por una persona. Cinco son los factores que determinan la complejidad del Reconocimiento del Habla:

A. *El Locutor*

Es quizás el aspecto que introduce mayor variabilidad en la forma de onda entrante, y por tanto requiere que el sistema de reconocimiento sea altamente robusto. Una persona no pronuncia siempre de la misma forma, debido a distintas situaciones físicas y psicológicas (es la llamada variabilidad intra-locutor). Existe además gran variedad entre distintos locutores (hombres, mujeres, niños), diferencias según la edad o la región de origen (variabilidad interlocutor). Es mucho más sencillo que un sistema funcione para un determinado locutor y que este lo haya entrenado previamente (se dice que el sistema es dependiente del locutor), a que un sistema funcione para cualquier locutor (sistema independiente del locutor).

B. *La forma de hablar*

Es el segundo factor que determina la complejidad de un reconocedor de habla. El hombre pronuncia las palabras de una forma continua, y debido a la inercia de los órganos articulatorios, que no pueden moverse instantáneamente, se producen efectos coarticulatorios. Ello, unido a las variaciones introducidas por la prosodia, hace que una palabra al principio de una frase sea diferente que cuando se dice en medio, o que sea diferente dependiendo de que es lo que le procede o le sigue. Un reconocedor es relativamente sencillo si sólo tiene que reconocer una palabra dicha de forma aislada (reconocedor de palabras aisladas) y es más complejo si debe reconocer las palabras de una frase, pero introduciendo una pausa entre cada dos de ellas (habla conectada). El sistema más complicado es aquel que debe funcionar reconociendo habla continua, que es la forma natural de hablar.

C. *El Vocabulario*

Se conoce por tal el número de palabras diferentes que debe reconocer el sistema. Mientras mayor es el número de palabras más difícil es el reconocedor, por dos motivos. El primero porque al aumentar el número de palabras es más fácil que aparezcan palabras parecidas entre sí, y el segundo porque el tiempo de tratamiento aumenta al aumentar el número de palabras con las que comparar. Una solución posible a este problema sería el utilizar unidades lingüísticas inferiores a la palabra (alófonos, sílabas,etc.) que en principio tienen un número limitado, e inferior al de posibles palabras. Sin embargo, la dificultad de reconocer estas unidades es aun mayor debido a que su duración es muy corta, la frontera entre dos unidades sucesivas es muy difícil de establecer y los efectos coarticulatorios son mucho más fuertes que entre palabras.

D. *La Gramática*

Es el conjunto de reglas que limita el número de combinaciones permitidas de las palabras del vocabulario. En general la existencia de una gramática en un reconocedor ayuda a mejorar la tasa de reconocimiento, al eliminar ambigüedades y puede ayudar a disminuir la necesidad de cálculo, al limitar el número de palabras en una determinada fase del reconocimiento ("perplejidad" de la gramática). En sistemas de palabras aisladas en los que no existe una gramática en el sentido estricto del término, se puede entender por tal el número de palabras a reconocer. Si, por ejemplo, el sistema debe reconocer un número telefónico urbano, la gramática de este sistema dice que el vocabulario son los diez dígitos, y debe reconocer un conjunto de siete dígitos, de forma que si el sistema reconoce más o menos, es que hay algún error.

E. *El Entorno físico*

Es una parte tan importante como las anteriores para definir el reconocedor. No es lo mismo un sistema que funciona en un ambiente poco ruidoso, como puede ser el despacho de un medico, o el que tiene que funcionar en un coche o en una fabrica. 0 por ejemplo, el que debe de funcionar a través de la línea telefónica, con la consiguiente reducción de banda o el que recibe la voz a través de un micrófono, que tiene mayor ancho de banda que la línea telefónica.

IV. EN QUE CONSISTE EL RECONOCIMIENTO DE VOZ

Podríamos afirmar que, genéricamente, el principal objetivo que el Reconocimiento de Habla persigue es proporcionar

una "apropiada" interacción hombre-máquina a través de órdenes habladas. Así, los resultados que esta tecnología proporcione deberán contrastarse con los derivados de otras alternativas como son: teclados, paneles, ratones, etc., en cuanto a si proporcionan un control de procesos de interacción hombre-máquina más o menos "apropiado". Las principales características que diferencian a los sistemas basados en Reconocimiento del Habla frente a otras alternativas son: la naturalidad que supone utilizar el habla en las operaciones de comando y control, y la precisión y robustez en la comunicación para diferentes usuarios y diferentes entornos. La primera de ellas debería representar la ventaja natural de los sistemas basados en la Tecnología del Habla. Aunque la experiencia nos ha enseñado que, si bien el habla es la forma natural de comunicación entre personas, en el diálogo hombre-máquina esto no parece obvio; piénsese, por ejemplo, en los diversos estudios que reflejan el elevado número de personas incapaces de responder frente a una máquina. Si bien es cierto que este tipo de rechazos va disminuyendo paulatinamente. Es la segunda de las características anteriores la que se muestra más crítica en las aplicaciones del Reconocimiento del Habla. El estado actual de la investigación en Reconocimiento del Habla nos muestra excelentes resultados de sistemas trabajando en entornos controlados de laboratorio. Sin embargo, una aplicación real de esta tecnología exige un funcionamiento en el mundo real donde el grado de dificultad de los problemas es un orden de magnitud mayor.

Bajo esa premisa de buscar una aplicación real, el modelo genérico de comunicación que el Reconocimiento del Habla propone para el diálogo hombre-máquina puede representarse, de forma simplificada, tal y como muestra el diagrama de la figura 1, para un caso de acceso a una base de datos.

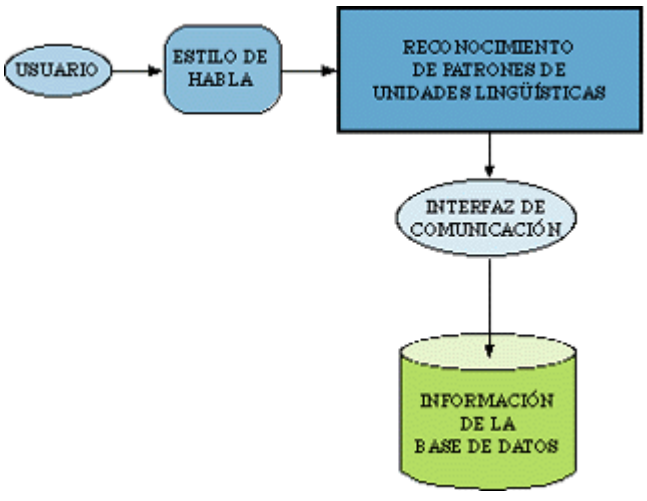

Figura 1. Modelo genérico de comunicación para Reconocimiento del Habla.

En este diagrama, el acceso a la información contenida en una base de datos comienza con la producción de un mensaje hablado por el usuario, pero utilizando una forma o estilo de habla restringido; por ejemplo, utilizando palabras de un vocabulario reducido pronunciadas de forma aislada (como los dígitos), frases tipo, etc. A partir de la señal de voz, un proceso de clasificación, basado en reconocimiento de patrones asociados a diferentes unidades lingüísticas (palabras, fonemas, sílabas, etc.), permite a una interfaz de comunicaciones extraer de la base de datos la información solicitada por el usuario.

Siguiendo el modelo de la figura 1 podemos presentar las principales áreas de trabajo que intervienen en el diseño y especificación de sistemas de Reconocimiento del Habla actuales. Estas áreas serían las siguientes:

Proceso de la señal de voz.

Técnicas de reconocimiento de patrones.

Diferentes estilos de habla.

Dependencia del locutor.

Vocabulario de reconocimiento.

Tarea de reconocimiento.

Bases de datos para entrenamiento y reconocimiento.

A. *proceso de la señal de voz*

La primera operación que debe realizar un reconocedor es procesar la señal de voz de entrada al sistema, con objeto de extraer la información acústica relevante para la tarea que debemos realizar. En este primer nivel del sistema son dos los interrogantes a resolver:

¿Qué rasgos o características extraer?
¿Qué efectos perturbadores pueden acompañar a la voz? y ¿cómo eliminarlos?
La respuesta a la primera cuestión ha venido precedida de un largo proceso de investigación sobre diferentes procedimientos de parametrización de la voz. Planteándose como solución actual más extendida una parametrización de la envolvente espectral que incluya consideraciones preceptúales a partir del funcionamiento del oído. Para reducir el número de parámetros posibles, la parametrización se combina con la utilización de técnicas discriminativas, seleccionándose el subconjunto con los parámetros más eficientes o distintivos [1].

En cuanto a la segunda de las preguntas planteadas, la presencia de efectos perturbadores en la señal de entrada, ha generado tres líneas de trabajo principales:

1) Detección robusta de voz: Apareciendo innumerables procedimientos de discriminación entre voz o ruido (silencio) para diferentes tipos de ruido [2].

2) Reducción de ruido: Distinguiéndose procedimientos que actúan directamente sobre la señal de voz y procedimientos que buscan compensar el efecto del ruido sobre la parametrización de la voz [3].

3) Cancelación de ecos: Incorporando técnicas de filtrado adaptativo que permitan al usuario comenzar a hablar mientras, desde el terminal remoto, se le está comunicando un mensaje que puede provocar un eco en la voz que entra al reconocedor [4].

B. *técnicas de reconocimiento de patrones*

El reconocimiento de patrones es la técnica más específica de todo sistema de reconocimiento. De ahí que muchos reconocedores se identifiquen a partir de la técnica de reconocimiento de patrones que incorporan. A partir de la representación paramétrica de la voz, este módulo realiza un proceso de clasificación utilizando una serie de patrones. Estos patrones se obtienen en una fase de entrenamiento del sistema y son representativos de un conjunto de unidades lingüísticas (palabras, sílabas, sonidos, fonemas). La peculiaridad más característica de este proceso, que marca su dificultad, es la variabilidad temporal que puede presentar una misma unidad lingüística al ser producida por diferentes modos y/o velocidades de habla. Así pues, las primeras técnicas de reconocimiento de patrones utilizadas fueron las basadas en un Alineamiento Temporal a través de algoritmos de Programación Dinámica, técnicas DTW [5]. Posteriormente se recurrió a la mayor flexibilidad que el modelado de procesos estocásticos permite para representar secuencias de duración variable. Concretamente la alternativa a las técnicas DTW fueron los Modelos Ocultos de Markov [6], (HMM), que pueden verse como una generalización de algoritmos DTW y han demostrado mejores prestaciones en multitud de sistemas de reconocimiento. También hay que mencionar que, recientemente, la potencia y excelentes capacidades de clasificación mostradas por las denominadas Redes Neuronales Artificiales (RN) las sitúa como posible alternativa frente a los HMM [7]. Hasta el momento las Redes Neuronales han permitido obtener los mejores resultados en Reconocimiento de Locutores, sin embargo en Reconocimiento del Habla encuentran como mayor dificultad la forma de afrontar la variabilidad temporal del habla.

Más adelante se explicaran con detenimiento estos métodos.

C. Modelado dependiente del estilo de habla

Se distinguen tres modos fundamentales de hablar frente a un sistema de reconocimiento:

Palabras aisladas

Supone que el usuario pronuncia una sola palabra o comando que el sistema deberá reconocer.

Habla conectada

El usuario pronuncia de forma fluida un mensaje utilizando un vocabulario muy restringido; el ejemplo más típico sería la pronunciación de un número telefónico.

Habla continua

Corresponde al modo más avanzado de funcionamiento de un reconocedor, y supone la pronunciación de frases de forma natural para un vocabulario amplio de palabras.

Además de los tres modos fundamentales anteriores, los reconocedores de voz tienen que afrontar, para un modelado robusto del habla, los tres aspectos siguientes:

1) Reconocimiento en contexto o "word spotting"

Técnica especialmente utilizada en reconocimiento de palabras aisladas, encaminada a detectar la presencia de palabras del vocabulario a reconocer en el contexto de otras palabras o pronunciaciones. La mayoría de las veces el contexto es resultado de la dificultad que encuentra el usuario para ceñirse a la pronunciación de una única palabra aislada. En otras ocasiones, el reconocimiento en contexto es la solución apropiada para robustecer el reconocimiento en ambientes acústicamente hostiles; por ejemplo, cuando la palabra que pronuncia el usuario viene acompañada de ruidos telefónicos, urbanos, etc. En cualquier caso, se trata de una técnica importante para robustecer los sistemas en aplicaciones reales.

2) Rechazo

Otro efecto de la presencia de sonidos indeseados (ruidos, sonidos o palabras fuera del vocabulario), es provocar el reconocimiento de palabras que realmente no han sido pronunciadas. Los procedimientos conocidos como técnicas de rechazo tienen como objetivo permitir incluir entre los resultados de reconocimiento la identificación de esos sonidos indeseados. Nos encontramos ante un problema de gran importancia de cara a la operatividad de un sistema de reconocimiento, que aún hoy por hoy no cuenta con una clara solución.

3) Múltiples candidatos

El proceso de reconocimiento de patrones que realiza un reconocedor se basa en identificar el patrón que ofrezca la puntuación más alta para decidir cuál es la mejor palabra o secuencia de palabras reconocida. Este proceso se basa en información exclusivamente acústica, sin tener en consideración otras posibles fuentes de conocimiento que podrían utilizarse para completar las puntuaciones de las diferentes palabras o secuencias candidatas. En la mayoría de los casos, la aplicación en que se encuentra el reconocedor es la que posee la información necesaria que permitiría seleccionar entre varias hipótesis de reconocimiento. Pensemos, por ejemplo, en una aplicación basada en el reconocimiento de números telefónicos; en esa situación, ante las dos hipótesis mejores de reconocimiento, una compuesta de cinco dígitos y otra de siete, la aplicación seleccionaría esta última independientemente de quién obtuviese la mayor puntuación "acústica" en el proceso de clasificación. Los procedimientos que permiten a un reconocedor disponer de la flexibilidad que supone manejar N hipótesis de reconocimiento se denominan N-best [8].

D. dependencia del locutor

El grado de dependencia del locutor define si el sistema incorpora patrones de unidades lingüísticas adaptados a un locutor determinado, y, por tanto, sólo funcionará correctamente para él, o si los patrones pretenden ser válidos para cualquier hablante. En el primer caso se habla de reconocimiento dependiente del locutor, mientras que en el segundo de reconocimiento independiente del locutor. A parte de las actividades específicas que se desarrollan para sistemas dependientes e independientes del locutor, existe un importante número de esfuerzos dirigidos a conseguir la

adaptación de un reconocedor a un locutor específico con la menor cantidad de voz posible [9].

E. *dependencia del vocabulario*

Las prestaciones de un reconocedor dependen fuertemente del tamaño y grado de dificultad del vocabulario. Es decir, del número de palabras que el sistema es capaz de reconocer, y de la mayor o menor dificultad de su reconocimiento en base a las relaciones de similitud fonética entre palabras. En la actualidad se diseñan sistemas tanto para vocabularios pequeños (menos de 50 palabras) y medios (entre 50 y 500 palabras), como para grandes vocabularios (más de 500 palabras), llegándose hasta 50.000 palabras para aplicaciones de dictado o acceso a bases de datos mediante lenguaje natural.

Otra importante dimensión, en relación con el vocabulario, es la que afecta a la distinción entre vocabularios fijos y flexibles. Una determinada aplicación, cuando esté reconociendo, siempre actuará sobre un vocabulario fijo. Pero en muchos casos ese vocabulario deberá variarse o actualizarse para eliminar y/o dar cabida a nuevas palabras. Tradicionalmente, una variación del vocabulario suponía comenzar un largo y costoso proceso de recogida de una nueva base de datos y re-entrenamiento de los patrones del sistema. En la actualidad hay diversas aproximaciones para conseguir un sistema con vocabulario flexible, que no necesite re-entrenarse para cada nuevo vocabulario [10].

F. gramáticas de reconocimiento

Según aumenta el número de palabras del vocabulario, el número de posibles combinaciones crece exponencialmente. Por tanto, se hace imprescindible la incorporación de restricciones, en cuanto al número de combinaciones válidas, según la tarea en que se inserte el sistema. Restricciones que suelen incorporarse en forma de gramáticas basadas en reglas sintácticas y/o semánticas destinadas a reducir el número de palabras susceptibles de ser reconocidas en cada momento. La medida utilizada para definir el grado de dificultad que supone una determinada tarea es la denominada perplejidad [11], de modo que un nivel de perplejidad bajo supone que en cada momento el número de posibles palabras candidatas es bajo, mientras que una perplejidad alta supone que ese número es alto, y consiguientemente el reconocimiento será más difícil.

V. TÉCNICAS DE DISEÑO

Se van a estudiar a continuación cuatro técnicas distintas que se utilizan o se han utilizado para el diseño de reconocedores de habla. De ahora en adelante se llamara "palabra" a la unidad básica en la que se base el reconocedor (en la realidad pueden ser sílabas, demisilabas, fenones, morfemas, palabras, conjuntos de palabras etc.). Las técnicas son:

Técnicas topológicas: Dynamic Time Warping (DTW), basado en el cálculo y comparación de distancias.

Técnicas probabilísticas: Modelos ocultos de Markov (HMM), que son modelos generativos de las palabras del vocabulario.

Redes neuronales.

Sistemas basados en el conocimiento: reconocedores por reglas o sistemas expertos.

En los cuatro casos se puede hablar de una fase de "entrenamiento" (cálculo de los patrones de referencia, cálculo de los parámetros de los modelos de Markov, entrenamiento de las redes neuronales o creación de estructuras de datos para los sistemas expertos) y de otra fase de "reconocimiento" propiamente dicho. Y también en los cuatro casos el primer proceso necesario es la "parametrización" o transformación de la forma de onda de la señal entrante en un conjunto de parámetros o características adecuadas a cada reconocedor.

A.Dynamic Time Warping

Los reconocedores de habla basados en técnicas de Dynamic Time Warping (DTW) han sido los primeros que han alcanzado un nivel de fiabilidad suficientemente alto como para dar lugar al desarrollo de productos comerciales.

Los sistemas de reconocimiento basados en DTW funcionan de la siguiente manera: Primero se parametriza la señal de voz a reconocer; para ello se divide en pequeñas ventanas de análisis (unos 20 mseg), y sobre cada una de esas ventanas se realiza un proceso de análisis que extrae un conjunto de parámetros (que pueden ser acústicos o coeficientes espectrales). Ese conjunto o vector de parámetros se puede ver como un punto en un espacio n-dimensional. El conjunto de todas las ventanas de análisis se convertirá así en una secuencia de puntos en ese espacio, y esa secuencia de puntos es lo que se llama "patrón" o "plantilla".

El sistema reconocedor dispone de un conjunto de patrones de "referencia" que se hayan calculado en la fase de entrenamiento, y que representan al conjunto de palabras del vocabulario que el sistema puede reconocer. De esta forma, una vez obtenida la plantilla de la palabra, la tarea del reconocedor consiste en compararla con todos los patrones de referencia que el sistema tiene, calculando la "distancia" que la separa de las referencias, y elegir como palabra reconocida aquella cuya plantilla de referencia de la menor distancia en la comparación.

Normalmente esas distancias se calcularían como la suma:

$$d_{XY} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} (x_{ij} - y_{ij})^2 \right)^{0,5}$$

Donde X es la plantilla de entrada, formada por m vectores de dimensión n, e Y es la referencia, también formada por m vectores de dimensión n.

El problema surge cuando X e Y tienen distinto número de vectores (lo cual se deberá a la distinta duración de la pronunciación de las palabras X e Y): ¿Qué hacer con los

vectores que sobran del patrón más largo?. Las técnicas de programación dinámica resuelven este problema: si X tiene m2 vectores e Y tiene m2 vectores, lo que se hace es "deformar" el eje de tiempos, estirándolo o encogiéndolo a voluntad para alinear ambos patrones de forma que vectores que representen sonidos iguales (o lo más parecidos posible) queden enfrentados a la hora de calcular las distancias. Así la distancia entre las dos plantillas se calcula siguiendo estos pasos:

1) Se calcula la matriz de distancias locales d(i,j) entre cada vector i del patrón de entrada X y cada vector j del de referencia Y, obteniendo una matriz de dimensiones [ml x m2].

2) Se calcula la matriz de distancias acumuladas g(i,j), utilizando las distancias locales d(i,j) según la formula recursiva:

3) **g(i,j) = d(i,j) t mintg(i-l,j), g(i- I ,j- I), g(i,j-l)**

4) Es decir, la distancia acumulada entre dos vectores es la suma entre su distancia local y la distancia acumulada mínima de los puntos vecinos anteriores en el tiempo.

5) La distancia total entre X e Y es la distancia acumulada entre los últimos vectores de ambas plantillas: g(m1,m2). La figura 2 muestra como podría quedar la alineación entre dos patrones de longitudes m1 y m2.

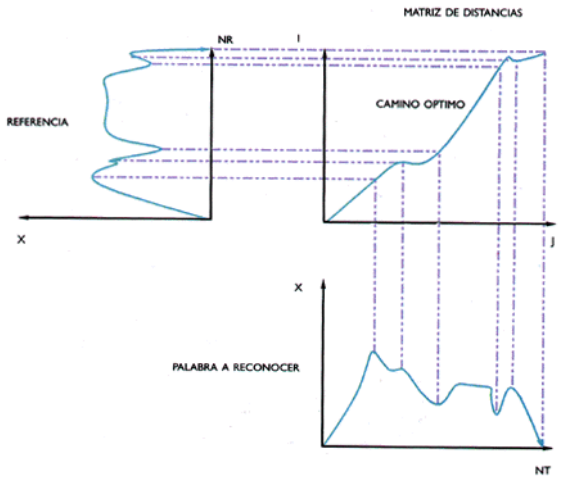

Figura 2: Alineamiento temporal entre la "Referencia" y la "Palabra a reconocer

El algoritmo que se acaba de describir es una versión muy simple df DTW.

Esta técnica ha sido la primera que ha permitido sacar productos a mercado, por las tasas de reconocimiento tan elevadas que produce (por encima del 98%, según la literatura). Hoy en día se ha abandonado, dejando paso a otras más modernas que, con tasas de error equivalentes, precisan menor volumen de cómputo en la tarea de re conocimiento, y menor necesidad de memoria.

B. Modelos ocultos de Markov

Otro enfoque alternativo al de medir distancias entre patrones (enfoque topográfico) es el de adoptar un modelo estadístico (paramétrico) para cada una de las palabras del vocabulario de reconocimiento, como son los modelos ocultos de Markov (HMM, del ingles 'Hidder Markov Models') [12].

Estos sistemas han sido posteriores en el tiempo, y hoy día la mayoría de los reconocedores en funcionamiento se basan en esta técnica estadística, ya que aunque sus prestaciones son similares a las de loa sistemas basados en DTW, requieren menos memoria física y ofrecen un mejor tiempo de respuesta. Tienen como contrapartida una fase de entrenamiento mucho más lenta y costosa, pero como esta tarea se realiza una única vez, y se lleva a cabo en los laboratorios. Es un precio que parece valer la pena pagar.

Un HMM se puede ver como una máquina de estados finitos en que el siguiente estado depende únicamente del estado actual, y asociado a cada transición entre estados se produce un vector de observaciones o parámetros (correspondiente a un punto del espacio n-dimensional del que se hablaba en el apartado anterior). Se puede así decir que un modelo de Markov lleva asociados dos procesos: uno oculto (no observable directamente) correspondiente a las transiciones entre estados, y otro observable (y directamente relacionado con el primero), cuyas realizaciones son los vectores de parámetros que se producen desde cada estado y que forman la plantilla a reconocer.

Para aplicar la teoría de los HMM en reconocimiento de voz, se representa cada palabra del vocabulario del reconocedor con un modelo generativo (que se calculara en la fase de entrenamiento) y posteriormente, se calcula la probabilidad de que la palabra a reconocer haya sido producida por cada uno de los modelos de la base de datos del reconocedor. Para ello, se asume que durante la pronunciación de una palabra, el aparato fonador puede adoptar sólo un número(finito de configuraciones articulatorias (o estados), y que desde cada uno de esos estados se producen uno o. varios vectores de observación (puntos de la plantilla), cuyas características espectrales dependerán (probabilisticamente) del estado en el que se hayan generado Así vista la generación de la palabra, las características espectrales de cada fragmento de señal dependen del estado activo en cada instante, y la evolución del espectro de la señal durante la pronunciación de una palabra depende de la ley de transición entre estados.

La representación más usual de un HMM es la utilizada para máquinas de estados finitos, es decir, conjuntos de nodos (que representar a los estados) y arcos (transiciones permitidas entre los estados). Un tipo de HMMs especialmente apropiado para reconocimiento de voz son los modelos "de izquierda a derecha"; modelos en los que una vez que se ha abandonado un estado, ya no se puede volver a él. La figura 3 representa un modelo con 'n' estados en el que desde cada estado sólo se permiten tres tipos de transición: al propio estado, al estado vecino y a dos estados más allá (este tipo de saltos que da recogido en una matriz de transiciones tridiagonal).

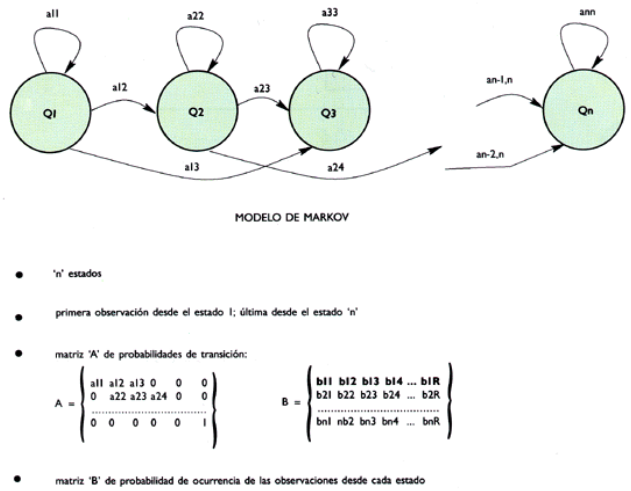

Figura 3:Modelo de Markov con 'n' estados

En cuanto a la generación de puntos de la plantilla, en estos modelos se asume que el primer vector de observaciones se produce desde el primer estado, y el último se emite desde el último estado. Recuérdese que la secuencia de estados es la parte oculta del modelo: se conocen los vectores de parámetros, pero no desde que estado se han producido.

1). definición formal de un hmm

Un modelo M viene determinado por los siguientes parámetros:

a) N- Número de estados del modelo.

b) Matriz de transiciones, de dimensión (N x N). Define la estructura del modelo: cada uno de sus elementos, aij, define la probabilidad de pasar del estado i al estado j. Normalmente A será bidiagonal o tridiagonal, significando que desde cada estado se pueden producir dos o tres tipos distintos de transición.

c) Conjunto de funciones de densidad de probabilidad (fdp) que modelan estadísticamente las observaciones producidas desde cada estado. Habrá pues tantas fdps como estados.

d) P- Vector de dimensión N. Cada uno de sus elementos, Pi indica la probabilidad de encontrarse inicialmente en el estado i. Para modelos de izquierda a derecha, Pl = I, y Pj=0 para los demás estados.

Como en el caso de DTW, la señal de voz viene representada por una plantilla o secuencia de vectores de características O=[Og, O2,...OT}, donde cada Oj es un conjunto de parámetros (coeficientes LPC, Cepstrum, log-area ratios...) que caracteriza la señal de voz en una ventana de tiempo centrada en t= i, y T es el número total de puntos de la plantilla. Los modelos HMM basados en este tipo d observaciones se llaman HMM continuos[26], y B será un conjunto de fdps continuas. Si, para simplificar las cosas, se hace pasar esa secuencia de observaciones O = {° I, O2,...OT} a través de un cuantificador vectorial (en que cada vector de parámetros Oi es codificado como u número entero [13]), la señal de voz quedara representada por una secuencia de centroides del cuantificador. Los HMMs que trabajan sobre este tipo de datos se conocen como HMM discretos, y B ser una matriz con tantas filas como estados tenga el modelo y tantas columnas como centroides tenga el codificador vectorial, en que cada elemento bjk es la probabilidad de que, estando en el estado i. s produzca el centroide k.

2). *reconocedor de palabras basado en hmms*

Una vez definido lo que es un modelo de Markov, se describe a continuación como se aplica a un problema real: el de reconocimiento de palabras (la metodología a usar sería la misma si se utilizasen otras unidades acústicas: fonemas, demisilabas, frases cortas, etc.).

El reconocedor dispondrá de un modelo por cada palabra del vocabulario de reconocimiento, y la estructura de esos modelos se define en la fase de diseño: el número de estados (N) se elige "a priori' según la complejidad que se pueda permitir y la calidad deseada. Valores típicos de N son entre 5 y 15 estados. Lo mismo ocurre con el tipo de transiciones: la matriz A tendrá sólo ciertas componentes distintas de cero, y su número es un parámetro de diseño. El tipo de funciones estadísticas que se utilizaran para modelar las probabilidades de observación de los puntos de la plantilla desde cada estado, también se fija antes de entrar en la fase de entrenamiento de los modelos. Suelen ser gaussianas multivariables, combinaciones lineales de gaussianas multivariables, funciones gamma, etc.

Una vez fija la estructura de los modelos se lanza la fase de entrenamiento, con el fin de calcular los valores óptimos de todos los parámetros que se han mencionado. Para ello, se usa un cierto número de repeticiones de cada palabra del vocabulario, que depende del tipo de reconocedor que se quiera construir (dependiente o independiente del locutor), de las prestaciones esperadas del sistema y del tipo de unidades que formen el vocabulario. Se puede decir que ese número de repeticiones varia entre 4 o 5 y unos cuantos centenares, lo que da idea del volumen de datos y de cálculos necesario. Del análisis de todas esas repeticiones saldrá el conjunto de parámetros que define cada modelo de Markov, y que formara la base de datos del reconocedor.

En los siguientes apartados, se explica como calcular los parámetros de cada modelo (entrenamiento) y como calcular la probabilidad P(0/M) de que una secuencia [Ot] de observaciones correspondientes a alguna palabra desconocida haya sido producida por cada uno de los modelos de la base de datos (reconocimiento propiamente dicho)[14].

3). *entrenamiento de un hmm*

Ya se ha dicho que un modelo M de Markov queda definido por tres matrices: A, B y P. Los modelos que se utilizan en el Reconocimiento del Habla (los denominados "de izquierda a derecha")[27] tienen un vector P fijo (= (1,0,0,...0)), por lo que no es preciso reestimar sus componentes.

Para simplificar las cosas, supóngase que cada repetición de una palabra produce una secuencia de vectores de características O(j)= {O1,O2,...0T}, y que se dispone de k pronunciaciones de cada palabra [O(1).. .O(k)}.

Entrenar el modelo es calcular los valores ajj y bj (0t) de ese modelo usando las k secuencias de observaciones O(1),O(2).. F0(k) correspondientes a las k repeticiones de la palabra a modelar, y de forma que la probabilidad de que el modelo calculado haya producido esas k secuencias sea máxima. El procedimiento que se sigue para entrena los modelos se indica en el diagrama de bloques de la figura 4: usando las k repeticiones de la palabra, se genera un modelo inicial segmentando uniformemente todas las plantillas entre los estados del modelo, y extrayendo unos estadísticos de esa primera segmentación se calculan los parámetros de un modelo inicial que será utilizado para una nueva segmentación, y así sucesivamente hasta que se considere que el modelo es suficientemente bueno.

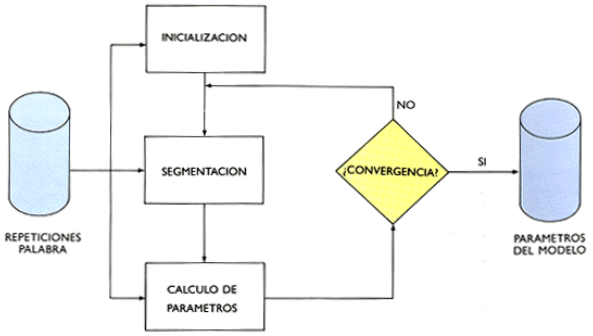

Figura 4: Entrenamiento HMMs

El algoritmo de Viterbi estima (usando el criterio de maximización a posteriori de P(0/M)) la secuencia más probable de estados durante la producción de la palabra, y la probabilidad final para esa secuencia de estados. Así, si se aplica Viterbi a cada una de las repeticiones de la palabra, se obtiene (usando las secuencias de estados) una partición de las observaciones, y se sabe desde que estado se ha producido cada una de ellas. Con estos datos, se recalculan los parámetros del modelo tal y como se indica en el siguiente ejemplo:

Supóngase que la secuencia de estados para la pronunciación de una palabra determinada es:

$$\underbrace{1,1,1\ldots1,}_{n_1}\ \underbrace{2,2\ldots2,}_{n_2}\ \underbrace{3,3\ldots3,\ \ldots\ldots,}_{n_3}\ \underbrace{N,N\ldots N}_{n_n}$$

Siendo:

nj = n° de veces que se ha visitado el estado i = n° de observaciones producidas desde el estado i.

Es decir, las nI primeras observaciones de la palabra se asignan al estado 1, las observaciones numeradas de nI + 1 a n1+n2 al estado 2, y así sucesivamente.

Una vez disponible esa segmentación para las k repeticiones de la palabra que se quiere modelar, las re-estimaciones de los parámetros del modelo correspondiente se hacen según las formulas:

$$a_{ii} = (N_i - k)/N_i$$

$$a_{i,i+1} = k/N_i$$

$$\text{Donde } N_i = \sum_{j=1}^{k} n_{ij}$$

Para la matriz B, en el caso de HMM continuos (funciones de densidad de probabilidad continuas: gaussianas, por ejemplo), los vectores de medias i y de varianzas i para cada estado i, se calculan promediando los vectores de observaciones Oj asignados al estado:

$$\mu_i = \left(\sum_{i=1}^{N_i} O_i \right) / N_i$$

$$\sigma_i = \left(\sum_{i=1}^{N_i} (O_i - \mu_i)^2 \right) / N_i$$

Esto es, después de la segmentación a cada estado se le asigna una partición del conjunto de las observaciones de las k repeticiones de la palabra; el valor medio de los vectores asignados al estado i será , i, y la varianza será i, calculada utilizando las diferencias entre el vector de medias y todas las observaciones asignadas al estado.

A continuación se describe el algoritmo de Viterbi.

4). Algoritmo de Viterbi

Este algoritmo [14], aplicado en reconocimiento de voz se utiliza para encontrar la secuencia de estados optima asociada a una secuencia de observaciones dada. Se basa, al igual que el algoritmo de Dynamic Time Warping en las técnicas de programación dinámica, y su formulación es:

Para encontrar la mejor secuencia de estados Q asociada a la secuencia de vectores de observación O dados por los vectores:

$$Q = \{q_1, q_2 \ldots q_t\}$$
$$O = \{O_1, O_2 \ldots O_t\}$$

Se define el conjunto de probabilidades acumuladas:

$$\delta_t(i) = \max_{q_i} P(q_1\ q_2..q_t = i, O_1\ O_2..O_t/M)$$

Que son las probabilidades de las secuencias óptimas de estados hasta el instante "t" y que terminan en el estado i. Se pueden expresar también como:

$$\delta_t(i) = \max \{P(q_1\ q_2..q_t = i/M) * P(O_1\ O_2..O_t/M)\}$$

$$\text{Para } t+1 \text{ se tiene } \delta_{t+1}(j) = \max_i \{\delta_t(i) * a_{ij}\} * b_j(O_{t+1})$$

El procedimiento completo para encontrar la mejor secuencia de estados es:

- Iniciación

$$\delta_1(1) = b_1(O_1)$$
$$\delta_i(1) = 0 \text{ para } i > 1.$$

- Finalización

$$P(O/M) = \delta_t(N)$$
$$q_t = N$$

- Obtención de la secuencia de estados

$$q_t = \tau_{t+1}(q_{t+1}),\ t=T-1,\ T-2,\ldots,1$$

5). etapa de reconocimiento

Dada una secuencia de observaciones O = (O1,O2,...Ot)l se calcula P(O/Mi), para I < i < Nw, siendo Nw el número de palabras del vocabulario, y se decide que O es la palabra representada por el modelo Mi que produjo la máxima probabilidad P(O/Mi). Esas probabilidades se calcularan también utilizando el algoritmo de Viterbi [15].

6). inclusión de modelos duracionales

Experimentalmente se ha comprobado la utilidad de modificar las probabilidades dadas por el algoritmo de Viterbi (P(O/M)) sumando otra cantidad directamente relacionada con la distribución temporal de la plantilla entre los estados del modelo. Esta modificación se hace a modo posproceso, en el sentido de que el algoritmo básico de reconocimiento no se ve afectado, únicamente la regla de decisión que determina la palabra del vocabulario elegida como palabra reconocida. La justificación teórica de la inclusión de los modelos temporales se puede encontrar en [16].

Durante la fase de entrenamiento de los modelos, y una vez que se ha determinado que esos modelos son suficientemente buenos, de la segmentación de todas las repeticiones de una misma palabra dada por Viterbi se pueden sacar estadísticas de la distribución temporal de las palabras entre los estados del modelo. Esas estadísticas (se modelara el tiempo transcurrido en cada estado como una gaussiana de media yj y desviación aj) se usaran en la etapa de reconocimiento para modificar las probabilidades dadas por Viterbi, en el sentido de favorecer al modelo que mejor se ajuste a la distribución temporal de la palabra a reconocer.

En la práctica se ha visto que la mejora que supone el uso de este tipo de posproceso es quizás insuficiente para justificar el incremento de carga computacional y de tiempo de ejecución que conlleva, en especial si el sistema reconocedor utiliza un bloque detector de extremos y trabaja en ambientes "no limpios".

C. Redes neuronales

Los modelos computacionales basados en redes neuronales surgieron hace ya relativamente bastante tiempo, pero se abandono su estudio por no disponer de algoritmos eficientes de entrenamiento. Ahora ya no existe esa dificultad, y se ha demostrado ampliamente su enorme potencia computacional.

Los sistemas de reconocimiento basados en redes neuronales pretenden, interconectando un conjunto de unidades de proceso (o neuronas) en paralelo (de forma similar que en la mente humana), obtener prestaciones de reconocimiento similares a las humanas, tanto en tiempo de respuesta como en tasa de error. Esa forma de interconexión de las unidades de proceso es especialmente útil en aplicaciones que requieren una gran potencia de cálculo para evaluar varias hipótesis en paralelo, como sucede en los problemas de reconocimiento de voz.

Las unidades de proceso pueden ser de varios tipos; las más simples (y utilizadas) disponen de varias entradas, y la salida es el resultado de aplicar alguna transformación no lineal a la combinación lineal de todas las entradas. Otro tipo de neuronas un poco más elaborado se caracteriza por disponer de memoria; en ellas la salida en cada momento depende de entradas anteriores en el tiempo.

La forma en que las neuronas se conectan entre si define la topología de la red, y se puede decir que el tipo de problemas que una red neuronal particular soluciona de forma eficiente, depende de la topología de la red, del tipo de neuronas que la forman, y la forma concreta en que se entrena la red.

Igual que se dijo para las técnicas anteriores, una red neural debe ser entrenada para resolver un tipo determinado de problemas. El algoritmo particular de entrenamiento dependerá de la estructura interna de las neuronas [17], pero, en cualquier caso, el entrenamiento se llevara a cabo a partir de una base de datos etiquetada, como sucedía con los modelos de Markov, y será un proceso iterativo en el que se modifican los parámetros de la red para que ante un conjunto determinado de estímulos (plantillas), produzca una respuesta determinada: la palabra del vocabulario representada por esas plantillas

La red neural que mejores resultados está dando hasta este momento en reconocimiento automático del habla es la denominada "perceptrón multicapa". La figura 5 muestra su topología: las neuronas se disponen por "capas"; hay una capa de entrada, que opera directamente sobre los vectores de observación o puntos de las plantillas, una capa de salida que apunta la palabra reconocida, y una o más capas intermedias. Cada capa está compuesta por varias unidades de proceso, que se conectan con la siguiente capa por una serie de enlaces a los que se da un cierto peso especifico wij.

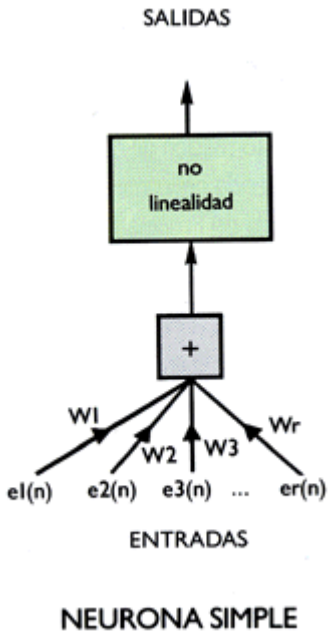

Figura 5: Neurona. Red Neuronal

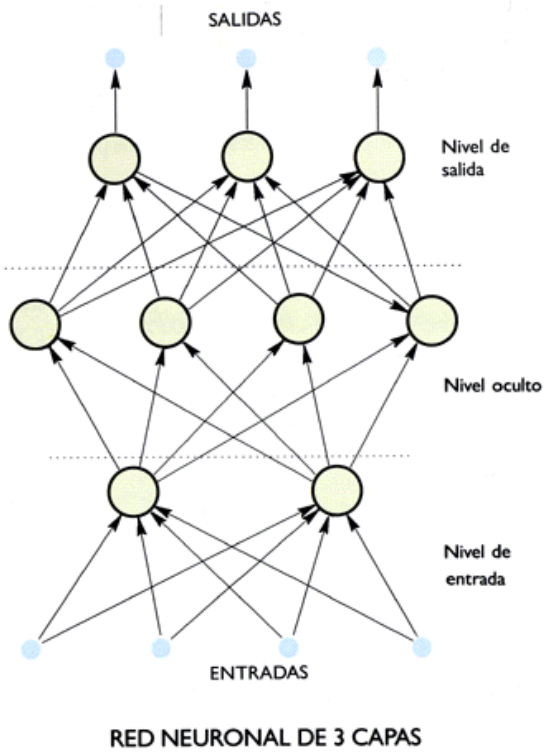

Figura 6: Red Neuronal

El conjunto de vectores de características entra en la capa de neuronas de entrada, y posteriormente es propagado a las capas siguientes. En cada célula de proceso se calcula la suma ponderada (por los pesos wij) de las señales de entrada, y posteriormente se procesa en la neurona con su sistema no lineal. Si el resultado de esta operación supera un cierto umbral, la neurona reacciona, transmitiendo señal a las neuronas siguientes de la capa superior.

En la fase de entrenamiento, dada una entrada conocida (p.ej. un conjunto de vectores que representa el dígito 1), la salida de la red es comparada con la salida esperada (y conocida de antemano), calculándose un error. Ese error se propaga hacia abajo, ajustándose de esta manera los pesos de las conexiones entre neuronas. Efectuándose este proceso varias veces se consigue que la red "aprenda" que respuesta debe dar para cada entrada en la fase de reconocimiento.

D. Reconocimiento basado en el conocimiento

Los métodos de reconocimiento descritos hasta ahora funcionan bastante bien cuando se trata de reconocer palabras aisladas. Cuando el sistema debe reconocer frases o habla continua, es necesario acudir a otras fuentes de conocimiento además de las puramente matemáticas y acústicas. Estas son por lo general reglas de tipo lingüístico, como se va a ver a continuación. Con este tipo de sistemas se llegara a tener no solo un reconocedor de habla sino un sistema de "comprensión" de habla.

La razón por la que a estos sistemas avanzados de reconocimiento se les llama Sistemas basados en el Conocimiento, se debe al uso de otras fuentes, otras disciplinas, otros conocimientos para llegar al entendimiento de la frase. En definitiva lo que se trata es que una máquina llegue a tener y utilizar los conocimientos que tiene una persona humana, para entender un mensaje.

1). Módulos básicos del sistema de reconocimiento

A continuación se describen los distintos niveles, o módulos básicos en que se podría subdividir un Sistema de Reconocimiento basado en el conocimiento.

a). Módulo de procesado acústico

En este módulo se extraen, a partir de la forma de onda de la señal de voz, un conjunto de parámetros representativos de la misma, que luego serán tratados en módulos posteriores. Para el cálculo de esos parámetros, se realiza un proceso de segmentación de la señal de entrada en pequeñas ventanas de análisis, y para cada una de las ventanas resultantes se calcula ese conjunto de parámetros, que pueden ser desde valores de la frecuencia fundamental, energía, densidad de cruces por cero y posición de los formantes, hasta otros parámetros que aporten información útil para comprender el sentido de la frase, como la variación de la frecuencia fundamental, la duración de los alófonos, etc[24].

b). Módulo de análisis fonético

Calcula, a partir de los parámetros obtenidos en el módulo anterior, la representación fonética más probable (o el conjunto de las más probables) correspondiente a la señal de voz. Esta transformación se basa en un proceso de etiquetado de los segmentos de análisis en que se divide la frase pronunciada, asignando a cada tramo de voz una

unidad lingüística abstracta, como pueden ser los alófonos. La ventaja de utilizar estas unidades para el siguiente tratamiento es que el número de datos a manejar es mucho menor, y además, debido a su naturaleza fonética, presentan una correspondencia bastante fuerte con la representación léxica.

Este proceso se denomina "categorización", y normalmente se realiza de acuerdo con un conjunto de reglas de producción. Por ejemplo:

IF < señal es cuasi_periódica .AND. frecuencia del primer formante baja.AND. frecuencia del segundo formante es alta>[18].

THEN <ventana de análisis se etiqueta como /i/>.

c). Módulo de análisis fonológico

El área de la fonología estudia la estructura o función de los sonidos dentro del lenguaje. El conocimiento fonológico permite la adaptación de los datos obtenidos en los niveles anteriores a una determinada lengua. Es necesario definir cuales son las unidades fonológicas que van a ser reconocidas en el Sistema de Reconocimiento: pueden ser alófonos, fonemas, difonemas, sílabas, palabras, etc. Estas unidades abstractas del lenguaje son estudiadas por separado y dentro de una secuencia para cada lengua en concreto.

Las reglas fonológicas aportan información de cómo varia la pronunciación de los fonemas, dependiendo del contexto. Con estas reglas se mejora o complementa la salida del Procesador Acústico-Fonético.

Para la realización de estas reglas, y un ajuste correcto de los parámetros, es necesario tener en cuenta la Prosodia de la frase. Los valores de los parámetros obtenidos en el análisis acústico-fonético ayudaran a determinar las sílabas tónicas o átonas, si la frase es enunciativa o interrogativa, etc.

d). Módulo de análisis morfológico

Es importante conocer, para cada lenguaje, las reglas de formación de las palabras a partir de los morfemas elementales. Esta es una de las facetas que estudia la morfología. Por ejemplo, hay combinaciones de sonidos o de letras que están permitidas en unos lenguajes y en otros no, por lo que es necesario conocer esas reglas de formación especificas. También hay reglas de formación de palabras a base de utilizar prefijos o sufijos. La disponibilidad de estas reglas, o incluso de un diccionario, ayudara a la determinación de palabras dentro de la cadena de unidades fonéticas que han salido del módulo acústico fonético.

Las reglas morfológicas ayudan también a la categorización gramatical de las palabras, lo que podrá ser usado por otros módulos.

e). Módulo de análisis sintáctico

La sintaxis estudia como combinar las palabras para construir frases de forma correcta en un determinado lenguaje. En cada idioma existe una serie de reglas de concatenación de palabras, constituyendo la Gramática del Lenguaje.

Un ejemplo de frase sintácticamente correcta sería: "El perro come la manzana". Un ejemplo de frase sintácticamente incorrecta sería. "El come manzana perro la".

Un sistema de reconocimiento que conozca y aplique las reglas de la sintaxis, ayudara bastante a decidir una secuencia lógica de palabras, y en caso de dudas entre los módulos anteriores, elegirá aquella que sintácticamente sea correcta.

Si un sistema debe reconocer una frase como "Los perros corren por el campo", ha podido tener dudas si "perros" va en singular o plural, dado que la terminación de la palabra es difícil de reconocer por el sistema y quizás el locutor no la ha dicho muy bien Sin embargo, si ha reconocido con bastante seguridad el artículo previo "Los", estará totalmente seguro que la palabra siguiente es "perros".

f). Módulo de análisis semántico

El conocimiento semántico está relacionado con cómo se encadenan las palabras para dar significado a una frase. Toma como partida el significado individual de las palabras, para deducir si una frase determinada tiene o no significado[19].

Una frase correcta desde el punto de vista semántico sería: "El pájaro está en el árbol". Sin embargo, la frase "El árbol está en el pájaro" es semánticamente incorrecta. Obsérvese que esta última frase es correcta sintácticamente.

En este módulo y los siguientes es donde empiezan los graves problemas de reconocimiento, ya que no se dispone aun de una forma eficiente de introducir este conocimiento en las máquinas.

Piénsese que hay muchas frases o palabras que tienen significado en un contexto y no lo tienen en otro, o lo tienen pero diferente. Esto es mucho más acusado en el lenguaje coloquial. Si por ejemplo se dice la frase "El pájaro estaba leyendo un libro", a nadie se le puede ocurrir que un pájaro pueda leer. Sin embargo si "El pájaro" es una denominación peyorativa de una persona, si que tendría sentido. ¿Cómo puede distinguir una máquina una opción de la otra?

g). Módulo de análisis pragmático

El nivel de conocimiento pragmático está relacionado con el contexto donde se están desarrollando las ideas.

Si se hubiera encabezado este artículo con la frase " La lluvia en Sevilla es una maravilla", nadie sabría la relación con el contenido del artículo. Sin embargo una frase como "El hablar con los ordenadores es un sueno que algún día se hará realidad", está relacionada con el tema del que luego se habla. La primera tiene un contenido sintáctico y semántico

correctos, pero pragmáticamente está fuera de contexto. No así la segunda. La utilización de este conocimiento está muy relacionado con el módulo de análisis semántico.

Puede darse el caso en que frases, sintácticamente mal formadas, tengan un contenido pragmático correcto. Esto es necesario tenerlo en cuenta, sobre todo en el contexto en que estamos de los Reconocedores de Habla, ya que sucede más veces en el lenguaje hablado que en el escrito[20].

h). Módulo de análisis del conocimiento del mundo

Este apartado incluye el conocimiento general que debe tener el usuario del lenguaje, con vistas, por ejemplo a mantener una conversación. Es necesario que se conozca el nivel de conocimientos del interlocutor en el tema de que se hable para que haya una transmisión de ideas.

Es totalmente ilógico que un premio Nóbel de medicina de una charla de bioquímica a un grupo de amas de casa utilizando un lenguaje totalmente técnico. Aunque todas las frases sean sintáctica, semántica y pragmáticamente correctas no habrá transmisión de ideas.

Dentro del tratamiento del lenguaje en los reconocedores de habla se puede utilizar para descartar hipótesis de palabras reconocidas, que por su complejidad técnica, estén fuera del alcance de la persona que está utilizando el reconocedor, o para incluirla si la situación es la contraria[25].

2). Estructura del sistema experto

La forma en que todas las fuentes de conocimiento que se han revisado se integran en el sistema reconocedor es un factor que influye decisivamente en la dificultad de implementación del sistema experto, y también en sus prestaciones finales.

La forma más simple de organizar todas esas estructuras de datos es de forma jerárquica (figura 7), dividiendo el trabajo entre varios bloques de proceso concatenados, cada uno de los cuales tiene como entrada la salida del procesador anterior en la cadena. Así, el procesador acústico-fonético analizando la forma de onda produce varias secuencias de fonemas, cada una de ellas correspondiente con un grado de probabilidad determinado a la trascripción fonética de la señal de entrada al sistema[21]. El procesador morfológico genera una red con las palabras más probables, y esa red pasa al procesador sintáctico, que la depura y recorta, dejando sólo las secuencias de palabras gramaticalmente correctas. El procesador semántico sigue limpiando esa red, eliminando las frases sin sentido. Por ultimo, y en el supuesto caso de que quede más de un candidato, será el procesador pragmático quien tome la ultima decisión.

Figura 7: Organización jerárquica

Ese sistema de organización permite el flujo de información en sólo un sentido, sin ningún tipo de realimentación que pueda aumenta la eficiencia del sistema. Se puede pensar en aprovechar, por ejemplo, la información del procesador pragmático (modulada por informaciones de tipo sintáctico y semántico) [22] para disminuir el número de posibilidades que los procesadores acústico-fonético y morfológico tienen que explorar. Este tipo de flujo "inverso" de información sin duda aumentara el tiempo de respuesta del sistema, así como la tasa de reconocimiento. La figura 8 muestra una estructura de interconexión que refleja esta idea.

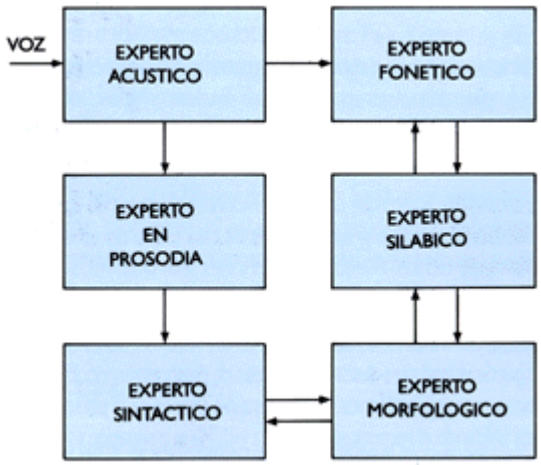

Figura 8: Sociedad de expertos

Otra organización diferente conectarla a todos los procesadores con cada uno de los demás utilizando el recurso de memoria compartida (por ejemplo). Esto queda reflejado en la figura 9. Es una estructura de más complicada implementación y mucho más versátil que ofrece más posibilidades de interacción que las anteriores. Sin embargo, parece que desborda un poco las necesidades de los sistemas expertos para reconocimiento tal y como están siendo concebidos hasta el momento[28].

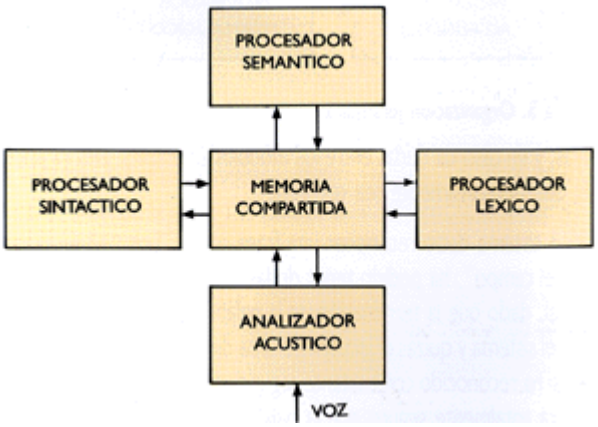

Figura 9: Organización con memoria compartida

VI. Conclusiones

En este artículo se resume los últimos avances obtenidos en los principales ámbitos del Reconocimiento del Habla, se ha realizado una presentación de la problemática, las

principales líneas de trabajo y las características particulares de los sistemas existentes. Se ha puesto especial énfasis en destacar los aspectos de innovación que incorporan los sistemas de Reconocimiento del Habla.

VII. BIBLIOGRAFÍA

[1] *Guarasa, M. Arquitecturas y métodos en sistemas de reconocimiento automático de habla de gran vocabulario universidad politécnica de madrid escuela técnica superior de ingenieros de telecomunicación, 2001*

[2] *Toledano, D. Segmentación y etiquetado fonéticos automáticos: un enfoque basado en modelos ocultos de markov y refinamiento posterior de las fronteras fonéticas*
señales, sistemas y radiocomunicaciones, escuela técnica superior de ingenieros de telecomunicación, universidad politécnica de madrid., 2000

[3] *Fernandez, D, Aportaciones a la mejora de los sistemas de reconocimiento universidade de vigo, 2001*

[4] *H. SAKOE and S. CHIBA: Dynamic Programming Optimization for Spoken Word Recognition. IEEE Trans. Acoust. Speech and Signal Proc., ASSP-26(1): 43-49 (1978).*

[5] *L. R. RAINER: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE 77(2), 257-286 (1989).*

[6] *GROIN and R. MAMMON: Introduction to the Special Issue on Neural Networks for Speech Processing. Speech and Audio Proc., vol. 1: 113-114 (1994).*

[7] *SONG and E. HUANG: A tree-trellis based fast search for finding the N best sentence hypotheses in continuous speech recognition. In Proc. CASS 91, pp. 537-540 (1991).*

[8] *L. RAINER and B-H JUAN: Fundamentals of Speech Recognition. Prentice Hall, pp. 449-450, New York (1993).*

[9] *E. BURKE, R. CARDIN, Y. NORMAN DIN, M. ROHM, J. WILSON: Application of Vector Quantized Hidden Markov Modeling to Telephone Network based Connected Digit Recognition. Proc. CASS (1994).*

[10] *R. M. SCHWARTZ, et al., Improved hidden Markov modeling of phonemes for continuous speech recognition. In Proc. CASS 84, vol. 3, paper 35.6, (1984).*

[11] *L. R. BAH, et al.: Acoustic Markov models used in the ANGORA speech recognition system. In Proc. CASS 88, vol. 1, pp. 497-500 (1988).*

[12] *X. HUANG, et al.: The SPHINX-II Speech Recognition System: An Overview. Tech. Report no. CMU-CS-92-112, CMU, Pittsburg (1992).*

[13] *HOW. HON: Vocabulary-Independent Speech Recognition: the VOICED System. Ph. D. Dissertation, CMU, Pittsburgh (1992).*

[14] *M.A. COHEN, et al.: The DECIPHER speech recognition system. In Proc. ICASSP-90 vol. 1, pp. 77-80 (1990).*

[15] *K. KITA, F. ANABATIC and H. SAITO: HMM continuous speech recognition using predictive OR parsing. In Proc. CASS 89, vol. 2, pp. 703-706 (1989).*

[16] *L. FISSURE, et al.: A word hypothesizer for a large vocabulary continuous speech understanding system. In Proc. CASS 89, vol. 1, pp. 453-456 (1989).*

[17] *R. NAY, et al.: Improvements in beam search for 10.000-word continuous speech recognition. In Proc. CASS 92, vol. I, pp. 9-12 (1992).*

[18] *J. G. WILSON and D. ROE: Applications of Speech Recognition Technology in Telecommunications. In Proc. ICSLP-94, pp. 667-670 (1994).*

[19] S. FRUIT: Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. IEEE Trans. Acoust. Speech and Signal Proc., ASSP-34(1): 52-59, Feb. 1986.

[20] *José A. Brito, JRH. Identificación de Señales Verbales en el Espacio de Fase Reconstruido Universidad de Los Andes, Postgrado en Computación,, 1999*

[21] *Nicolas Pecan, DOFF. Hams and OWE Neural Network for Continuous Speech Recognition 2001*

[22] *Jordá Adén, a.C.*
Análisis de la Segmentación Automática de Fonemas para la Síntesis de Voz.
2001

[23] *Ismael Cortázar Múgica, AMARC. Últimos desarrollos en tecnologías de voz y del lenguaje 2002*

[24] *Eduardo Clemente, a.C. Entrenamiento y Evaluación de reconocedores de Voz de Propósito General basados en Redes Neuronales feedforward y Modelos Ocultos de Harkov TALOTA-SENTÍA, 1999 , 15*

[25] *Taylor, J.F.K.R.S.K.P. an automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces.*

[26] Ahuactzin, I.K.N.A.A. Aplicación de Tecnología de Voz en la Enseñanza del Español *Universidad de las Américas- Puebla.*, 2001

[27] J.L. Gauvain, L.L. conversational telephone speech recognition *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2003 , 212-215

[28] Ries, K. hmm and neural network based speech act detection 1999