Capítulo 2 .- Redes Neuronales en reconocimiento de locutor.

Autores : Carlos Vivaracho Pascual[*], Luis Alonso Romero[**]
[*] Dpto de Informática. Universidad de Valladolid. España
[**] Dpto, de Informática y Automática. Universidad de Salamanca. España

2.1- Introducción

Buscando una mayor claridad en la exposición, se ha optado por un enfoque ascendente. Empezaremos centrando el tema de forma que queden claros todos los conceptos relacionados con el reconocimiento del locutor, así como sus aplicaciones prácticas más relevantes. A continuación, se hará un breve repaso al ñestado del arteò, cuyo objetivo es mostrar al lector las diferentes estrategias de clasificación empleadas, comentando el grado de utilización actual de cada una de ellas: se pretende, fundamentalmente, mostrar el estado actual de utilización de las redes neuronales, frente a otro tipo de modelos de clasificadores. Para terminar se desarrollará con un poco más de profundidad el tema que nos interesa: *las redes neuronales en el reconocimiento del locutor*, de forma que el lector pueda tener una idea completa de cual ha sido y es la aplicación de las redes neuronales en el reconocimiento del locutor. Se expondrán algunos resultados, de forma que se pueda comparar, con las reservas oportunas que se indicarán, el rendimiento tanto de los distintos tipos de redes neuronales entre si, como de estas frente a otros modelos usados en el reconocimiento del locutor.

Cualquier sistema de reconocimiento de locutor consta de dos partes fundamentales: extracción de características de la señal de voz representativas del locutor, clasificación del o de los vectores de características extraídos. Como se puede observar del enfoque indicado en el párrafo anterior, nos vamos a centrar en la segunda parte: clasificación, obviando todo lo referente a la etapa de extracción de características por considerar que carece de interés con respecto al tema que nos ocupa.

2.2.- Definición y Aplicaciones

El reconocimiento automático del locutor (RAL) o reconocimiento automático del hablante es una área de investigación y desarrollo de aplicaciones de gran importancia actualmente. Tan antigua como el reconocimiento automático del habla, a tenido, sin embargo, una atención, y como consecuencia, un desarrollo menor. Ha sido en estos últimos años cuando, a la luz de los importantes campos de aplicación que han surgido, el desarrollo y los esfuerzos investigadores han sido mayores, incentivados, además, por el creciente interés del mundo de la empresa. Esto ha hecho que hayan aparecido las primeras aplicaciones comerciales de aplicación más o menos restringida. Antes de continuar con un desarrollo más en profundidad del tema, vamos a intentar definir el problema y enumerar algunos de los más importantes campos de aplicación del RAL.

En el RAL se persigue el reconocimiento de las personas mediante su voz, sin que sea necesaria la intervención de un operador humano. Como es evidente, el ordenador será el encargado de sustituir al operador humano en la tarea de reconocimiento. Es fácil ver, que el RAL se enmarca dentro de un área de trabajo más amplia: el reconocimiento de personas mediante parámetros biológicos cuyo valor es diferente en cada persona, como por ejemplo el iris, la forma de la cara o las huellas dactilares, etc. Son los que se denominan parámetros biométricos.

Dentro del RAL podemos distinguir dos tareas diferenciadas:

* **Verificación Automática del Locutor** (VAL). El objetivo es verificar la identidad reclamada por el locutor, o sea, tenemos un individuo que dice ser alguien y una muestra de su voz, la tarea a realizar es ver si ambas coinciden o no; la respuesta del sistema será, por lo tanto, binaria: identidad aceptada o rechazada. Hay diversas formas de medir la efectividad de este tipo de sistemas, una de las más utilizadas es la denominada tasa de equierror (normalmente conocida por las siglas inglesas EER): es el error del sistema cuando el umbral decisión (si salida sistema < umbral la identidad

reclamada por el cliente es rechazada, en caso contrario será aceptada) es tal que el porcentaje de falsas aceptaciones es igual al de falsos rechazos (fig. 2.1).

* **Identificación Automática del Locutor** (IAL). Aquí el objetivo es, dada una muestra de voz, señalar, dentro de un grupo de personas, la identidad de su ñpropietarioò. Hablamos de IAL de Grupo Cerrado si el locutor desconocido en con certeza uno de los del grupo, y de IAL de Grupo Abierto si existe la posibilidad de que el locutor pueda ser alguien ajeno a ese grupo de personas. En el primer caso la respuesta del sistema será siempre una identidad, mientras que en el segundo existe la posibilidad de la respuesta hablante rechazado al no pertenecer al grupo de personas a identificar.

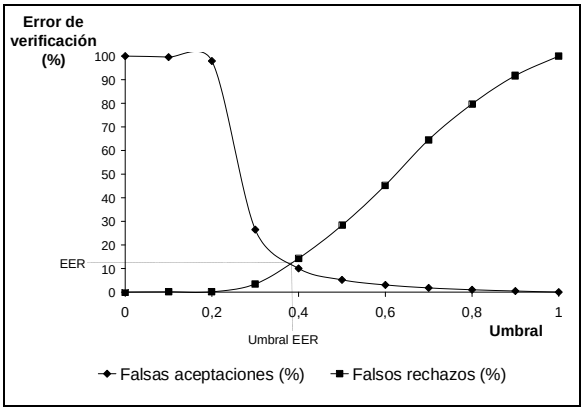

Figura 2.1. Curvas de error falsas aceptaciones (FA) falsos rechazos (FR). FA: la entrada a la red pertenece a un impostor y es falsamente aceptada. FR: la entrada a la red pertenece al cliente y es falsamente rechazada.

Según el contenido de la muestra de voz empleada tanto en la VAL, como en la IAL, podemos distinguir los siguientes dos casos extremos:

- **VAL o IAL Dependiente de Texto** (VALDT o IALDT), cuando el contenido de la muestra de voz es fijo y conocido de antemano por el ordenador (típicamente una palabra o frase clave).

- **VAL o IAL Independiente del Texto** (VALIT o IALIT), cuando o la muestra de voz es aleatoriamente escogida por el hablante, de forma que el sistema desconoce a priori su contenido.

Cada opción tiene sus ventajas e inconvenientes. Así, por ejemplo, el rendimiento de los sistemas que utilizan reconocimiento dependiente de texto es mayor, sin embargo son más vulnerables cuando la clave es conocida por personas ajenas a su propietario. El riesgo de la clave es evitado usando independencia del texto, pero el rendimiento de los sistemas que utilizan esta opción es menor, y, además, el peligro radica en que cualquier grabación de la voz de una persona puede ser utilizada para un acceso no permitido. Como alternativa surge una tercera vía, intermedia entre las dos anteriores:

* ñ**Text Promptedò** (VALTP o IALTP), cuya traducción podría ser ñtexto apuntadoò. Es el sistema el que escoge el contenido de la muestra de voz. Esta elección puede ser hecha dentro de un conjunto reducido de palabras o frases, o sin ningún tipo de restricción en el texto a pronunciar por el hablante. Como en el reconocimiento dependiente del texto, el sistema conoce a priori el contenido de la muestra de voz, por lo que los modelos usados en el reconocimiento pueden ser más aproximados mejorando éste. Al mismo tiempo, se evita la violación de la seguridad del sistema tanto por la perdida del secreto de la clave (ya que esta no es fija), como por grabaciones de la voz del hablante, ya que es imposible conocer a priori el texto que pedirá el sistema.

Definido el problema a tratar, vamos a hacer un breve repaso a algunas de las aplicaciones prácticas más relevantes del reconocimiento del locutor (Boves 1998):

* **Arresto domiciliario**: comprobar vía telefónica que el arrestado está en su domicilio, verificando su identidad mediante la voz. Esto evita la necesidad de una comprobación in situ.
* **Control de acceso** a estancias, a sistemas, a recursos, é : es quizás la aplicación más antigua.
* **Seguridad en la red**: es cada vez más frecuente el acceso a recursos remotos vía ordenador. Verificar la identidad del usuario en este campo se hace verdaderamente difícil. El uso de la voz para tal fin aparece como una opción viable.
* **Acceso a servicios financieros** sin necesidad de personarse en el banco.
* **Servicios telefónicos**: van desde la mejora en la seguridad en el uso del servicio de tarjeta telefónica, la prevención del fraude por robo del numero en telefonía móvil, o al control del acceso a cualquier servicio personalizado, por ejemplo, buzones de voz.

Para la mayoría de estas aplicaciones podemos encontrar en el mercado productos comerciales ya desarrollados. Aunque, quizás, la fiabilidad de la voz como parámetro biométrico es inferior, en estos momentos, al de otros como pudieran ser huellas dactilares o iris, el hecho de que en numerosas ocasiones sólo contemos con la voz del usuario para comprobar su identidad, augura un futuro prometedor al campo de investigación del reconocimiento del locutor.

2.3.- Estrategias de clasificación: evolución histórica y estado actual

Los sistemas de reconocimiento del locutor corren paralelos al os de reconocimiento del habla en los comienzos, de forma que las metodologías empleadas eran una adaptación de las usadas en estos últimos. Así, por reseñar las más relevantes, podemos encontrar sistemas basados en:

* **Parámetros Estadísticos (Long-Term-Statistics)**: Es de las primeras aproximaciones utilizadas. La serie temporal de vectores de características extraídos de la señal de voz es representada mediante parámetros estadísticos como su media y/o su varianza. Usado fundamentalmente en reconocimiento del locutor independiente del texto, podemos encontrar referencias de su uso ya en los años 70, por ejemplo, reseñar los trabajos de Furui (1972) y Markel (1977).

* **Alineamiento Temporal Dinámico (Dinamic Time Warping)**: Su aplicación más habitual es en sistemas de reconocimiento del locutor dependiente del texto. El locutor es representado mediante un conjunto de patrones, correspondientes cada una de las palabras o frases clave. La decisión sobre el reconocimiento o no del locutor es realizada a partir de la distancia, medida aplicando el alineamiento temporal dinámico, entre la muestra de voz del locutor, y los patrones de referencia. Un ejemplo típico de este tipo de sistemas se puede observar en la figura 2, donde se esquematiza el sistema planteado por Furui en 1981, para la verificación del locutor dependiente del texto.

* **Modelos Ocultos de Markov (HMM)**: Empleados en reconocimiento del locutor tanto dependiente, como independiente del texto, variando el tipo de modelo empleado en cada caso: izquierda a derecha en el dependiente del texto, y ergódico en el independiente del texto. Con respecto a este segundo caso, reseñar los trabajos de Poritz (1982), proponiendo un modelo ergódico de 5 estados, donde cada estado representaba una de las 5 amplias categorías en que dividió los fonemas. Con respecto al caso del reconocimiento del locutor dependiente del texto, indicar los trabajos, posteriores a las primeros referencias de uso de HMMs, de Zheng y Yuan (1988), entre otros, demostrando la superioridad de los sistemas basados en HMM, frente a los más tradicionales basados en DTW.

* **Cuantización vectorial (VQ)**: Dada su característica de no poder modelar la información temporal incluida en la secuencia de vectores de características extraída de la muestra de voz, su uso se plantea fundamentalmente en el reconocimiento del locutor independiente del texto. Modelo muy utilizado en los primeros sistemas (por ej., Li y Wrench 1983, Shikano 1985), ha sido muy estudiado en comparación con otros modelos como los HMMs, demostrando un mejor comportamiento cuando el número de muestras de voz de referencia es pequeño, ahora, bien, cuando éste aumenta, su rendimiento decrece frente a los HMMs (Matsui y Furui 1992).

La evolución de los estudios centrados en el tema, llevó a una profundización en las líneas de trabajo presentadas, teniendo como consecuencia el auge de las basadas en HMM, en detrimento de las otras, al tiempo que aparecieron nuevos enfoques, entre los que podemos destacar los sistemas basados en Modelos de Mezcla de Gaussianas (Gaussian Mixture Model, GMM) y Redes Neuronales Artificiales (ANN). Con respecto a los primeros, la técnica fue propuesta por Rose y Reynolds (1990), sus trabajos posteriores la han asentado y demostrado su efectividad. Cuando hablamos de redes neuronales en el reconocimiento del locutor estamos englobando, como veremos, un amplio espectro de sistemas, todos caracterizados por utilizar en alguna de sus partes un modelo conexionista. No vamos en este punto a profundizar más en el tema, ya que será objeto de estudio en el siguiente apartado, tan solo indicar, siguiendo la reseña histórica de esta sección, que las primeras referencias sobre el uso de redes neuronales en el reconocimiento del locutor quizás sean los trabajos de Oglesby y Mason (1990) donde usaban un MLP por hablante, y de Bennani y otros (1990) donde el modelo escogido fue un LVQ (Learning Vector Quantization). La situación actual presenta un claro predominio de las líneas de investigación y desarrollo basadas en modelos parámetricos, como los HMM y los GMM, usándose los primeros, fundamentalmente, en el reconocimiento del locutor dependiente del texto y texto pedido, mientras los segundos presentan claras ventajas en el caso del reconocimiento del locutor independiente del texto.

Antes de pasar al siguiente apartado creemos conveniente comentar el párrafo anterior. Que las redes neuronales son, actualmente, una metodología poco usada, es un hecho fácilmente constatable sin más que echar un vistazo a los ñproceedingsò de los 2 últimos congresos que sobre el tema tuvieron lugar en Martigny (Suiza), 1994 y en Avignon (Francia), 1998: mientras las referencias al uso de redes neuronales en el primero es prácticamente comparable al del resto de técnicas, incluso con sesiones y tutoriales especializados, en el segundo son muy escasas. Es difícil buscar razones objetivas, entendiendo como tales estudios comparativos profundos, que expliquen esta tendencia, en nuestra opinión son más razones de índole práctico o de posición de los grupos involucrados en el tema.

Nuestra experiencia, basada en un estudio comparativo entre HMMs y ANNs (concretamente MLP), sobre una misma base experimental (igual corpus, iguales pruebas), demuestran un mejor comportamiento de las ANNs, sobre todo en el reconocimiento del locutor independiente del texto, teniendo la ventaja de ser modelos con un menor número de parámetros. El resumen de los resultados obtenido puede observarse en la tabla 2. Aunque de alcance limitado, debido al reducido número de pruebas realizadas (el estudio era una primera aproximación), los resultados nos parecen alentadores, y a falta de un trabajo comparativo más profundo, en vías de realización, nuestra opinión es que la redes neuronales pueden tener su opción, todavía, en el reconocimiento del locutor.

2.4.- Las redes neuronales en el reconocimiento del locutor

Son numerosos los trabajos realizados empleando distintos modelos de redes neuronales y/o comparando el rendimiento de estas entre sí, ya que el ámbito de aplicación, con respecto a las distintas opciones de estudio que presenta el reconocimiento del locutor (aptdo. 2), es total, es decir, han sido utilizadas tanto en verificación, como en identificación, pudiendo ser estas tanto dependientes, como independientes de texto.

Generalmente relacionadas con la etapa de clasificación, el papel jugado dentro de los sistemas de reconocimiento del locutor es más amplio, y aunque en menor medida, también son utilizadas en la etapa de extracción de características de la señal de voz. En estos casos, han sido usadas para realizar transformaciones no lineales sobre el vector de características, buscando espacios de representación que mejoren la eficacia de la etapa de clasificación.

Centrándonos en esta etapa, indicar que las estrategias de clasificación en las que intervienen, de una u otra forma, redes neuronales son múltiples. Como se reflejó al referenciar los primeros trabajos sobre el tema en el apartado anterior, lo más clásico es utilizar para la clasificación de la señal de voz un determinado tipo de red neuronal: MLP, RBF, é Son numerosos los estudios basados en este tipo de sistemas, pudiendo diferenciar, fundamentalmente, dos enfoques distintos: en el más habitual las redes

tienen un papel discriminante, es decir, su salida proporciona directamente información acerca de la pertenencia o no de una muestra de voz a un determinado hablante. El otro enfoque consiste en usar las redes como predictores, pudiéndosele considerar como una extensión no lineal de los modelos autoregresivos: la entrada a la red es la secuencia de vectores de características x_{t-i}, é , x_t, siendo la salida la predicción del vector x_{t+1}, la información acerca del propietario de la señal de voz se obtiene de la comparación entre el vector predicción y el vector real.

Casi simultáneo con el estudio de los sistemas que acabamos de indicar, aparecen otros que buscan mejorar la etapa de clasificación, usando sistemas más complejos. Una gran línea de trabajo en este sentido, es la basada en el uso de modelos híbridos HMM-ANN, de forma que el resultado es un clasificador que usa lo mejor de cada uno para corregir las limitaciones de ambos. Dentro de esta línea podemos incluir, aunque no sean exactamente sistemas híbridos, aquellos trabajos que tras obtener mediante un modelo HMM la probabilidad a priori $P(x/\lambda)$: dado el modelo λ, que la muestra de voz x pertenezca al modelo, utilizan ésta como entrada a una red neuronal para obtener la probabilidad a posteriori $P(\lambda/x)$: dada la muestra x, que pertenezca al modelo λ, siendo esta probabilidad la salida final del sistema. La otra gran línea de trabajo es la basada en el uso de sistemas modulares, de forma que la decisión se basa en integrar la salida de múltiples clasificadores, pudiendo ser estos de igual naturaleza (por ej., mezcla de expertos), o de distinta naturaleza (por ej., reconocimiento de personas mediante la integración de reconocedores basados en distintos parámetros biométricos o integración de los resultados de distintos clasificadores ante una misma muestra de entrada). En este caso de sistemas modulares, la etapa de clasificación la podemos dividir en dos partes: respuesta individual de cada clasificador, integración de todas estas para lograr una salida única del sistema; pues bien, las ANNs pueden intervenir tanto en una tarea, como en la otra.

Como se puede observar, el campo de estudio es amplio. Buscando una mayor claridad en la exposición, hemos decidido dividir ésta atendiendo a un doble criterio de papel de la red neuronal dentro del sistema de reconocimiento del locutor, y estrategia de clasificación seguida:

* **Redes neuronales en la etapa de clasificación: sistemas simples**. Las redes forman parte de la etapa de clasificación. Aquí analizaremos sistemas en los que la clasificación se realiza mediante un único módulo, bien compuesto por un determinado tipo de red neuronal, o bien por un modelos híbrido HMM-ANN.

* **Redes neuronales en la etapa de clasificación: sistemas modulares**. Nuevamente las redes forman parte de la etapa de clasificación. Incluiremos en este apartado referencias tanto al papel jugado en la parte de clasificación, como en el de integración.

* **Otras aplicaciones.**

2.4.1.- Redes neuronales en la etapa de clasificación: sistemas simples

2.4.1.1.- Redes no recurrentes

Cuatro son los modelos principalmente usados: Perceptron Multicapa (MLP), red neuronal de Funciones de Base Radial (RBFNN), Learning Vector Quantization (LVQ) y Mapa autoorganizado de Kohonen (SOM), aunque como se verá, la aplicación de este último es peuqeña. Vamos a estudiar cada uno por separado, incluyendo las referencias oportunas a los estudios comparativos con el resto.

MLP

Aunque en los estudios realizados a tal fin, el rendimiento comparativo con otros modelos, por ejemplo RBF, no haya sido superior, si se puede considerar el modelo más frecuentemente utilizado y de manera más versátil. Dada esta amplitud de estudio, vamos a dividir la utilización del MLP en el reconocimiento del locutor en tres grandes grupos, atendiendo a su función:

* **Como clasificador**. La salida de la red es directamente una estimación de la probabilidad de pertenencia de la entrada a una determinada clase.
* **Como predictor**. Dada una secuencia de vectores de características correspondientes a los instantes de tiempo t-n, t-n-1, é , t-1, la red trata de predecir el valor del vector de características en el instante siguiente, o sea, t.
* **Como transformada**. Se busca una transformación sobre los datos de entrada que sea característica del hablante.

MLP como clasificador

Una de las tareas en las que ha demostrado un alto rendimiento es como generador de **funciones discriminantes**: la salida se puede considerar como la probabilidad de que la muestra de voz pertenezca o no a una determinada clase (hablante, en nuestro caso). Aunque conceptualmente el papel de la red es similar en todos ello, vamos a ver algunos sistemas propuestos por diversos autores. En sus primeros trabajos, Oglesby y Mason (1990) abordan la tarea de identificación dependiente de texto, basada en una secuencia de 10 dígitos. La población de hablantes era de 10, entrenando para cada uno una red neuronal distinta, mediante una única muestra de la secuencia de 10 dígitos. Una vez concluido el entrenamiento la asignación muestra de voz-hablante se realiza siguiendo el criterio de red con valor de salida más alto. El sistema se probó con 4 muestras de la secuencia de 10 dígitos por hablante (distintas, obviamente, a las de entrenamiento). Se experimento con distintas arquitecturas, obteniendo los mejores resultados: 8% de error en la identificación, con una de tres capas, siendo el tamaño de la oculta de 128 neuronas. Las características usadas fueron 10 coeficientes cepstrales, obtenidos a partir de los coeficientes de predicción lineal (LPCC). Los autores compararon el sistema propuesto con otro basado en VQ, obteniendo para éste un rendimiento similar al mostrado cuando el número de centroides es de 64.

Vivaracho (1994) une el poder discrimante del MLP, con la compactación que, en cuanto al número de vectores de características a procesar, supone el uso de parámetros estadísticos para representar la muestra de voz. Concretamente, se utiliza la media por banda de frecuencia, calculadas estas mediante escala de Mel sobre el espectro de frecuencia, como entrada a la red. La perdida de información temporal que supone el promedio, se ve compensada al utilizar muestras de voz de corta duración, como son los dígitos, para el reconocimiento. Se realizaron pruebas de verificación tanto dependiente, como independiente de texto, sobre una población de 21 hablantes. En verificación dependiente de texto, se entrena una red por dígito y hablante, usando 5 muestras del dígito para tal fin. La red es probada con las otras 5 muestras del dígito de que se dispone. En verificación independiente del texto, se entrena una red por hablante, usando para ello las 10 muestras de 8 de los 10 dígitos. La red se prueba con las 10 muestras de los 2 dígitos que no fueron usados en entrenamiento. Tanto en un caso como en otro, la red es entrenada para dar salida 1 si la muestra pertenece al hablante, y 0 en caso contrario, usando como criterio de convergencia que la diferencia entre la salida deseada y la real para todas las muestras de entrenamiento está por debajo de un determinado valor umbral, fijado a 0.01. En la tabla 1 se pueden observar los resultados en el caso más favorable: red con tres capas, 32 neuronas en la de entrada y 4 en la oculta, comparando estos con los de un sistema basado en HMMs (trabajo realizado Silva, Vivaracho, Alonso y Cardeñoso en 1998). Un dato a tener en cuenta en la comparación es que el número de parámetros de la red es muy inferior al necesario en HMM.

Hennebert y Petrovska (1998), abordan la tarea de verificación del locutor mediante texto pedido (text prompted). Como primera aproximación, en su trabajo estudian el comportamiento de un sistema de verificación dependiente de texto basado en fonemas individuales, de forma que esto les permite analizar el rendimiento de distintos grupos de estos en la tarea propuesta. Entrena una red por hablante y fonema con una arquitectura, nuevamente, de tres capas, con 20 neuronas en la oculta y, a diferencia de los otros dos trabajos presentados, 2 en la de salida: una para la clase cliente (hablante para el que se entrena la red) y la otra para la clase impostor (todo hablante distinto al cliente): en entrenamiento las salidas deseadas se fijan a 1 y 0 respectivamente si la muestra pertenece al hablante cliente, y a 0 y 1 en caso contrario. El coeficiente de aprendizaje η es actualizado tras cada época de entrenamiento, según el siguiente criterio:

* $\eta_{i+1} = \eta_i/2$ si el error medido sobre un conjunto de validación (independiente del de entrenamiento) se ha incrementado en la época i con respecto al obtenido en la época i-1.

* $\eta_{i+1} = \eta_i$ si el error anteriormente indicado ha decrecido

El hecho de que se incremente el error indica que la red está sobreentrenando, para evitar esto la actualización de pesos de la red es descartada si ocurre la situación planteada en el primer punto anterior. El entrenamiento se finaliza cuando η caiga por debajo de un valor umbral prefijado. En prueba la decisión (hablante aceptado/rechazado) se toma atendiendo al criterio de la clase de la neurona con el valor de salida mayor. Las muestras de voz utilizadas en los experimentos pertenecen a la ñHER Swiss German telephone speech databaseò, y fueron tomadas vía teléfono en una única sesión. Una vez aislados los fonemas, las muestras correspondientes a cada uno de estos son divididas en ventanas: porciones de 30 ms. de señal, extrayendo de cada una un vector de características compuesto por 12 LPCC. Para probar la influencia de la información contextual en el reconocimiento del locutor, experimenta el sistema con y sin esta información, o sea:

* Sin información contextual: tanto en entrenamiento como en prueba la entrada a la red es el vector de características extraído de cada una de las ventanas en que ha sido dividida la señal de voz correspondiente.
* Con información contextual: ahora la entrada a la red incluye además del vector de características de cada ventana, los de j ventanas anteriores y posteriores. Los valores de j probados fueron 1,2 y 3.

El rendimiento del sistema es mayor cuando se incluye información contextual, obteniéndose la mejor relación resultados/tamaño de la red para j=1. En este caso, el valor de la tasa de equierror varía entre el 9.7%, en el mejor de los casos, para fonemas nasales y el 22.5%, en el peor de los casos, para sonidos líquidos.

Para concluir con trabajos que usan MLP como funciones discriminantes, referenciar el realizado por Labanova y Raev (1998). En su estudio presentan un enfoque original en cuanto a los datos de entrada a la red, proporcionando a la salida de esta un significado también original. Su objetivo es la verificación dependiente de texto vía teléfono. Como parámetros representativos del locutor parten de los formantes, pero no de los valores estáticos que estos adquieren en cada una de las ventanas en que se divide la señal (concretamente, de un tamaño de 256 muestras, con un solapamiento de 40 entre dos consecutivas), sino de la evolución de estos en porciones de como máximo 6 ventanas, con la característica añadida de que esa evolución debe ser suave (este concepto es implantado estableciendo un umbral en la diferencia entre un formante de una ventana y la siguiente); a las zonas así obtenidas las denominan CFR (Chain of the Frame). El reconocimiento de una muestra de voz de origen desconocido se hace como sigue: se realiza un alineamiento entre CFRs de la muestra de entrenamiento y CFRs de la desconocida, sobre estos se establecen una serie de medidas (concretamente 8 distintas), cuyos valores serán las entradas a la red. La salida de esta será, por los tanto, la probabilidad de que las dos muestras confronten, o sea, la probabilidad de que los hablantes comparados sean el mismo o no. Dado que varias de las entradas a la red se pueden considerar como medidas de similitud entre dos vectores, el papel de está también puede ser considerado como de integrador no lineal de todas estas medidas para obtener una respuesta única, algo similar a lo que veremos en el siguiente apartado. La arquitectura seleccionada vuelve a ser de tres capas, con 16 neuronas en la oculta y 1 en la de salida. Se realizaron pruebas con 25 hablantes clientes y 100 impostores, obteniendo una tasa de equierror de 2.9 %, para claves basadas en sentencias de duración de 3 a 5 segundos.

MLP como predictor

La eficacia del MLP como discriminante ha sido de sobra demostrada, sin embargo, es conveniente recordar que para que las superficies de separación entre clases puedan ser perfectamente definidas, es necesario presentarle a la red, en entrenamiento, un conjunto de ejemplos de cada una lo suficientemente representativo. Particularizando al caso del reconocimiento del locutor, esto, en principio, es totalmente factible en identificación, donde tenemos muestras de todas las clases a diferenciar. Sin embargo, la tarea de verificación es diferente: hemos de diferenciar entre un hablante (llamemosle cliente) y el resto (impostores), donde el ñrestoò es cualquier hablante. Esto plantea dos cuestiones a tener en cuenta: primero, la de poseer un conjunto de muestras de hablantes impostores, y, segundo, como escoger éste

para que sea suficientemente representativo. Los enfoques que veremos a continuación obvian este doble problema: incluso en verificación, la red es entrenada con tan solo las muestras del hablante cliente.

Aparte de su capacidad como clasificadores, algunos autores han explotado el potencial del MLP como generador de funciones, en este caso para la **predicción** de series temporales. Los sistemas presentados pueden ser considerados como una extensión no lineal de los modelos autorregresivos lineales (AR) estudiados por autores como Montacié y Le Floch (1992) o Bimbot y otros (1992).

Uno de los principales autores que han profundizado en esta campo ha sido Hattori. En el trabajo presentado en ICASSP´92, estudia 2 enfoques distintos para un sistema de identificación independiente del texto. En el primero de ellos entrena una red por hablante de forma que sea capaz de predecir una ventana de voz a partir de las dos anteriores. Durante la fase de prueba se mide la diferencia entre los valores pronosticados y los reales, asignando la muestra de voz a aquella red (hablante, por tanto) con un error de predicción menor. El enfoque presentado está suponiendo que el proceso que se intenta modelar es estacionario, pero en realidad no lo es. Para tratar este problema se plantea un sistema con M estados, donde cada estado es un modelo predictivo (MLP) diferente, especializado en una determinada parte de la muestra de voz del locutor. Este sistema puede ser considerado como un HMM, con una probabilidad de transición fija entre estados, y donde cada uno de estos ha sido sustituido por un MLP; estamos hablando de un sistema híbrido HMM-MLP, que será tratado en el apartado correspondiente .

El mismo autor presentó en el congreso especializado en el reconocimiento del locutor, celebrado en Martigny (Suiza) (1994), un sistema para la verificación del locutor independiente del texto, que es una modificación del de un solo estado anteriormente descrito (le llamaremos sistema base). La modificación afecta a la salida de la red, a la que aplica lo que denomina una normalización del error de predicción. Para la realización de esta normalización parte de la siguiente hipótesis: el error de predicción aumenta debido a dos factores, uno son las propias limitaciones de la red en su proceso de aprendizaje, le llama factor inherente, y el otro son aquellas características que no pueden ser aprendidas de los datos de entrenamiento, denominándole factor no aprendido. Ejemplos de factores inherentes son el ruido no causal: la red asume causalidad entre el ruido y las señales de voz vecinas, y el hecho de usar para la predicción un número de ventanas inferior al de la causalidad en la señal de voz. El factor no aprendido es la diferencia entre los datos de entrenamiento y los de prueba. Este último factor es difícilmente evitable, por lo que la normalización pretende paliar de alguna forma el primero. La idea es dividir el error de predicción obtenido del sistema base E_b, entre un valor E_i que, careciendo de la información acerca del hablante, tenga el mismo error inherente. Este segundo error se logra mediante una red con la misma arquitectura y entrenada de la misma forma que la del hablante cliente, pero con muestras de N hablantes.

El sistema fue probado para 12 hablantes hombres (con similares características), usando 150 palabras para entrenamiento y el resto para prueba. La muestras de voz fueron obtenidas mediante micrófono en una habitación preparada. Como características se utilizaron 10 coeficientes cepstrales obtenidos a partir de la división del espectro de frecuencia mediante escala de Mel (MFCC). La arquitectura de las redes neuronales utilizadas es tres capas, con 20 neuronas en la de entrada, 20 en la oculta y 10 en la de salida. Se entrenó una red por hablante, siendo la usada para la normalización común a todos. Los resultados muestran una mejora considerable con la introducción de la normalización, así por ejemplo, al usar un conjunto de 10 palabras para probar el sistema, pasamos de un EER del 41% sin normalización al 2.7% con ella. Comparando el rendimiento del modelo propuesto con otros, tales como VQ y HMM, se llega a resultados similares, pero con un menor número de parámetros en el sistema basado en redes neuronales (casi la mitad).

Otros autores que han utilizado modelos predictivos en el reconocimiento del locutor son Artieres y otros (1993) realizando una comparación extensiva entre distintos modelos, Levin (1990) proponiendo una modificación del modelo de M estados presentado por Hattori: el modelo de control oculto (Hidden control model, HCNN), que ha demostrado su efectividad en determinadas situaciones, pero con un mayor número de neuronas en la capa oculta, en general, que con el modelo inicial, y Sorensen y Hartman (1993) proponiendo un HCNN auto estructurado, donde el número de neuronas va incrementándose durante el entrenamiento para adaptarse a la dificultad del problema; pruebas realizadas en identificación muestran unos resultados peores que con HCNN.

MLP como función transformada

En este caso se busca realizar transformaciones sobre los vectores de características extraídos de la señal que sean características de cada hablante.

Un enfoque similar al planteado en el caso del MLP como predictor, es el presentado por Lastrucci, Gori y Soda (1994), que proponen el uso de MLPs como **autoasociadores**: se fuerza a la red a reproducir la entrada en la salida. Es una situación similar a la planteada en el uso del MLP para la compresión de datos. En prueba la decisión está basada en la distancia euclídea entre la entrada y la salida a la red: si la distancia medida supera un umbral prefijado es hablante es rechazado, siendo aceptado en caso contrario. Este enfoque es aplicado a un sistema de verificación del locutor dependiente de texto, basado en fonemas (concretamente, solo se probó con 2: /ae/ y /aa/ y de forma separada), utilizando como base de datos la DARPA-TIMIT. Como características extraídas de la señal de voz usan 20 LPCC. Con las secuencias de vectores de este tipo extraídos de las muestras de aprendizaje se entrena una red por hablante, con una arquitectura de tres capas, con 20 neuronas en la de entrada, 6 en la oculta y, obviamente, 20 en la de salida. Para la verificación de la identidad del locutor prueban inicialmente dos criterios: basar la decisión en la distancia d_i medida sobre una única ventana de voz, o utilizar un conjunto N de estas, concretamente de 5. En este segundo caso, la distancia es promediada de la siguiente forma:

$$D = \frac{1}{N} \sum_{i=1}^{N} (e^{C \bullet d_i} + K)$$

Donde C y K son constates para optimizar el efecto de la función exponencial. La tasa de equierror obtenida al usar el fonema /ae/ es de 7.9 al usar una sola ventana, bajando al 6% al usar 5. Los resultados son peores con el otro fonema. Tanto con un criterio como otro, aunque más con el primero, la decisión está basada en un tamaño de señal de voz excesivamente pequeño, por lo que introducen un tercer criterio establecido, en este caso, a nivel de fonema: agrupan todas las ventanas de voz correspondientes al fonema en conjuntos de 5, y sobre cada uno calculan D; es suficiente con que una distancia promediada D esté por debajo del umbral para que el fonema sea asignado al hablante bajo prueba. Los resultados del sistema mejoran considerablemente con este último criterio. Por ejemplo, en las pruebas realizadas con el fonema /ae/, y con un umbral de decisión fijo, se pasa de unos porcentajes del 46% en falsas aceptaciones y 1.3% en falsos rechazos con el segundo de los criterios indicados a nivel de ventana, a unos porcentajes del 0% y el 3.5% respectivamente. Similares mejoras se observan con el fonema /aa/.

Otra referencia a un sistema similar al anterior lo podemos encontrar en el trabajo de Gong y Haton (1994). Los autores prueban tres configuraciones distintas para entradas I_a y las correspondientes salidas deseadas I_b de la red:

* I_a e I_b pertenecen a distintos fonemas, pero pronunciados por el mismo hablante. Se busca que la red capture lo que comparten pronunciaciones de distintos sonidos por un mismo hablante.
* I_a e I_b son secuencias de vectores de características del mismo fonema pero emitidas I_a por el hablante cliente e I_b por un hablante de referencia común a todos los hablantes clientes. Se supone que la transformación obtenida es particular a cada hablante cliente.
* I_a es igual a I_b, o sea, pertenecen al mismo hablante y al mismo sonido.

Los autores señalan que esta última configuración, que es la similar a la presentada por Lastrucci, Gori y Soda, es la que mejores resultados da. El sistema es probado tanto en verificación como en identificación dependiente del texto: la palabra clave es, tan sólo, un fonema. Para ambas tareas se emplean las mismas redes (tres capas, con 4 neuronas en la oculta) entrenadas para cada hablante con 4 muestras de cada uno de los 4 fonemas usados como posibles palabras clave. Como referencia, en identificación, asignando la muestra al hablante con menor distancia acumulada entre entrada y salida de la red, el error obtenido es del 3.9%, para una población de 72 hablantes. Resaltar, tal y como indican los autores, la corta duración de las muestra de voz usadas tanto para entrenar las redes, como para el reconocimiento del locutor.

En una línea de trabajo similar a la del autores anteriores, Narendranath y otros (1994), muestran la eficacia del MLP para transformar características de voz de un hablante en características de voz de otro hablante. Concretamente, entrenan y prueban MLPs para transformar formantes vocálicos de un hablante, en fomantes vocálicos de otro.

Hermansky y Malayath (1998) presentan en su trabajo un enfoque totalmente diferente. Parten del hecho de que cualquier muestra de voz contiene 3 tipos de información diferentes: información lingüística (qué se dice), información específica del hablante (quién lo dice) e información acerca del canal y/o el entorno en que se produce la comunicación. En el reconocimiento del locutor el interés de centra en aislar la segunda de las tres. Para intentar modelar esa información, los autores proponen un método con los siguientes tres pasos:

* Extraer de la señal de voz del hablante vectores de características, I, con información exclusivamente lingüística.
* Extraer de la misma muestra de voz, otro conjunto de vectores de características, D, pero ahora con información tanto lingüística como del hablante.
* Estimar una función transformación, M, entre D e I, tal que la distancia entre D y M(I) sea mínima. Una vez obtenida M, en prueba, esa distancia será la utilizada para la clasificación de la muestra de voz de origen desconocido.

Notar que como D e I tienen el mismo contenido lingüístico, M transportará la información que está presente en I y no en D, es decir, la información específica del hablante. Para la obtención del vector D, dependiente del hablante, los autores prueban tres tipos de características: coeficiente cepstrales PLP de orden alto (concretamente de orden 14), coeficientes de predicción lineal (LPCs) y energía por banda de frecuencia usando escala de Mel. Para la obtención del vector I, independiente del hablante, los autores proponen dos tipos de características: coeficiente cepstrales PLP de orden bajo (concretamente de 7º orden) y las extraídas de la aplicación de una técnica que denominan ñOriented Principal Component Analysisò, OPCA (Malayath, Hermansky y Kain 1997). Para implementar la función transformación se propone el uso de un MLP, con una única capa oculta con 30 neuronas. Comparando el sistema propuesto con el más clásico basado en mezcla de gaussianas (GMM) en verificación independiente del texto, bajo las mismas condiciones experimentales, los autores llegan a la conclusión de que el rendimiento de ambos es similar, pero con la ventaja del modelo propuesto de emplear menos parámetros (750 frente a los 4864 del GMM). También se observa una ligera pero consistente mejora en los resultados del nuevo modelo, cuando se aumenta la duración de la muestra de voz empleada para probar el sistema. En cuanto a las características usadas para D e I, indicar la ventaja de la energía por banda de frecuencia y las basadas en OPCA, respectivamente.

RBFNN

Han sido generalmente utilizadas como una alternativa al Perceptron multicapa en clasificación directa de patrones: tienen similares propiedades discrimantes. Así por ejemplo, Oglesby y Mason en el trabajo presentado en ICASSPô90 (ya comentado en la parte referida al Perceptron), identifican como principales problemas del MLP su excesivo tiempo de entrenamiento y el excesivo tamaño que alcanza la red cuando la complejidad del problema aumenta. Para intentar aliviar estos problemas los mismos autores proponen en su trabajo del 1991 (ICASSô91) el uso de RBFs. Las condiciones experimentales son similares a las ya presentadas anteriormente para el MLP, en cuanto a que se entrena una red por hablante. El sistema fue probado en verificación, para 40 hablantes, obteniendo, según los autores, un rendimiento mayor que con sistemas basados en MLP y VQ. El mejor resultado, para una secuencia de prueba compuesta por 4 dígitos, y con una red con 384 neuronas en la capa oculta, es de un 8% en falsos rechazos y un 1% falsas aceptaciones.

Fredrickson y Tarassenko (1994) plantean un sistema de identificación dependiente del texto, basado en letras aisladas del alfabeto. Como funciones de activación de base radial de la primera capa de la

RBFNN, se usan gaussianas de matriz de covarianza diagonal; la salida de cada una de estas neuronas será:

$$\Phi(\|\vec{x}_p - \vec{c}_j\|) = \exp(-\sum_{i=1}^{N_I} \frac{(x_i - c_{ji})^2}{2\sigma_{ji}^2})$$

Donde $\vec{x}_p$ es el p-ésimo vector de entrada a la red, N_I es el número de componentes de estos vectores, $\vec{c}_j$ es el centroide correspondiente a la neurona j de la capa oculta y σ_j es su matriz de covarianza diagonal. La segunda capa, la de salida de la red, realiza una aplicación lineal sobre las salidas de la capa anterior:

$$y_{kp} = w_{k0} + \sum_{j=1}^{N_C} w_{kj} \Phi(\|\vec{x}_p - \vec{c}_j\|)$$

Donde y_{kp} es el k-esimo nodo de salida de la red, w_k es su vector de pesos y N_C es número de neuronas en la capa anterior.

Los primeros experimentos son un estudio comparativo entre MLP y RBF, obteniendo una mayor eficacia con este segundo tipo de redes. La base de datos utilizada para estas pruebas está compuesta por emisiones aisladas de las 10 primeras letras del alfabeto, realizadas por 6 hablantes distintos; la grabación se realizó mediante un micrófono. La señal de voz correspondiente a cada letra es dividida en ventanas de 256 muestras, extrayendo de cada ventana 8 coeficientes cepstrales vía FFT; mediante una normalización temporal lineal el número de vectores de características se reduce a 8 para todas las muestras de voz: todas serán representadas, por tanto, mediante un vector de características final de 64 componentes (entradas a la red). Tanto en el caso MLP, como RBFNN, para realizar la identificación se entrena una única red con 6 salidas, 1 para cada hablante, usando para ello 10 muestras de cada letra del alfabeto (600 patrones de entrenamiento). El mejor resultado en el caso MLP se obtiene con una red con 24 neuronas en la capa oculta y es de un 29.5% de error. Para el caso RBF el peor de los resultados referenciados es de un 25.8% de error, para 60 neuronas en la capa oculta, que baja a un 21% para 300, límite máximo impuesto siguiendo la norma de que para que la red sea capaz de generalizar, el número de neuronas en la capa oculta tiene que ser menor que el número de patrones de entrenamiento; los autores fijaron el límite en la mitad de esta cantidad. La primera capa de la RBFNN fue entrenada mediante el algoritmo de los k-medios (k-means), y los pesos de la segunda mediante la técnica de la inversión de la matriz.

A raíz del resultado del estudio comparativo, los autores deciden centrar sus esfuerzos en un estudio más profundo del rendimiento de las RBFNN en la tarea propuesta. Tras un análisis del comportamiento individual de cada letra por separado, realizan la identificación usando conjuntos de tamaño r variable de estas. La salida final es el producto de las salidas individuales para cada letra: estas están siendo interpretadas como una probabilidad; el índice de la salida con un valor final mayor, identifica al hablante. Para una población de 50 hablantes (50 neuronas de salida en la red), usando 1000 patrones de entrenamiento (2 ejemplares por letra, 10 letras, para cada hablante), 500 patrones de prueba (1 ejemplar por letra para cada hablante) y con una red con 300 neuronas (centroides) en la capa oculta, los resultados obtenidos varían entre un 23.8% de error para r=1, 0.84% para r=5 y 0% para r=10, usando las 10 letras con mejores resultados, y unos errores de 31.38%, 2.83% y 0.2% respectivamente, usando 10 letras escogidas aleatoriamente.

LVQ y SOM

El LVQ es, quizás, junto al MLP uno de los primeros modelos conexionistas empleados en el reconocimiento del locutor. Surge como un refinamiento de una técnica más clásica como es el VQ. También es conveniente comentar que no ha tenido un uso muy extendido.

Una de las primeras referencias de utilización la encontramos en Bennani, Fogelman y Gallinari (1990). Más concretamente, en su trabajo intentan comparar el rendimiento de dos parametrizaciones diferentes de la señal de voz: 12 coeficientes LPC (Linear Predictive Coding) y 8 coeficientes MFCC

(Mel Frequency Cepstral Coding), usando para ello un sistema basado en LVQ. La tarea abordada en la identificación dependiente de texto, mediante un sentencia en francés de duración 2.5s., y para una población de 10 hablantes. El conjunto de vectores de características extraídos de la señal de voz correspondiente a una sentencia es representada mediante su media y el primer autovector de la matriz de covarianza: viendo el conjunto de vectores como una nube de puntos en el espacio de características, la media representa su posición, y el autovector su forma. La base de datos contiene 10 repeticiones de la sentencia usada en la identificación, de las cuales 9 son usadas para entrenar el modelo LVQ y la restante para prueba; se hacen 10 experimentos distintos variando el contenido del conjunto de entrenamiento y prueba (técnica ñleave-one-outò). El LVQ es inicializado mediante un ñk-meansò con K=2 o 3, para cada hablante. Los autores muestran unos resultados superiores con la parametrización MFCC, alcanzando tasas de error en la identificación inferiores al 3%.

Muy poco se ha escrito sobre la utilización de los mapas autoorganizados de Kohonen en el reconocimiento del locutor. Una de esas pocas referencias la podemos encontrar en el trabajo realizado por Anderson y Patterson (1994), que utilizan una red de Kohonen para inicializar un LVQ. El objetivo del trabajo indicado es comparar el rendimiento de parámetros auditivos, basados en los modelos de Patterson (Auditory Image Model, AIM) y de Payton (PAM), frente a parámetros más clásicos en el tratamiento de la voz como son los cepstrales, concretamente, MFCC. La tarea abordada es la identificación dependiente de texto, mediante sonidos vocálicos extraídos de sentencias de la base de datos TIMIT, y para una población de 37 hablantes. Entrenan una red, o mejor, un modelo LVQ por hablante, de la siguiente forma: con vocales extraídas de 7 sentencias correspondientes a una misma persona, se entrena un SOM de 8x8 neuronas, con disminución tanto del parámetro vecindad, como del de aprendizaje. Una vez entrenada se etiqueta cada nodo, asignándole a aquel fonema cuyas muestras más veces le hayan activado, para, a continuación, mediante un LVQ (concretamente LVQ3) realizar un ajuste más fino de los centroides: el resultado final es un modelo con 64 centroides. El proceso de identificación del hablante al que pertenece una determinada sentencia se realiza extrayendo las muestras correspondientes a las vocales, y midiendo la distorsión D_s media mínima para cada modelo entrenado s:

$$D_s = \frac{1}{N}\sum_{i=1}^{N} \min_{j \in k} ||x_i - m_{sj}||^2$$

Donde N es el número de vectores de características x_i, y el índice sobre los centroides m_{sj} es k=1, é , 64. La muestra de voz se asigna al hablante cuyo modelo tenga una distorsión menor. Las pruebas se realizaron con una sentencia por hablante (37 pruebas en total), obteniendo unos errores de verificación del 6% para los parámetros MFCC, del 9% para los basados en AIM y del 33% para los basados en PAM.

En el apartado de sistemas modulares veremos otra referencia de uso de los SOMs. En ese caso el SOM puede ser visto como un primer nivel en la etapa de clasificación o como una transformación del espacio de características.

2.4.1.2.- Redes recurrentes

Nuevamente nos encontramos con un tipo de redes muy poco utilizadas. Dentro de esos pocos trabajos presentados, referenciar el realizado por Tsoi, Shrimpton, Watson y Back (1994). Estos autores prueban el rendimiento de un modelo recurrente: arquitectura Frasconi-Gori-Soda (AFGS), comparando el rendimiento de éste frente al de un MLP en la misma tarea. La arquitectura indicada es una extensión del modelo más simple de Jordan-Elman, y pertenece a la clase de modelos recurrentes denominados ñlocally recurrent globally feedforwadò (LRGF), que se caracterizan porque la recurrencia es local a la neurona. Sobre un modelo de neurona clásico, modelo de McCulloch-Pitt, la recurrencia se puede establecer en tres puntos distintos:

* En las conexiones de entrada a la neurona: se establece una realimentación en cada una de las conexiones sinápticas.

* Sobre la salida de activación a(t): a las entradas a la neurona se añaden valores anteriores de salidas de activación de dicha neurona, o sea, se produce una realimentación (con los retardos correspondientes) de la salida de activación de la neurona a la entrada.
* Sobre la salida de la neurona y(t): es similar al caso anterior, pero ahora la realimentación se realiza después de haber aplicado la función no lineal sobre la salida de activación.

La AFGS se corresponde a este último caso, y en ella la salida de la neurona en un instante t será:

$$y(t)= f\left(\sum_{j=1}^{m} k_j y(t-j)+ \sum_{i=0}^{n} w_i x_i(t) \right)$$

Donde f va a ser una función sigmoide.

Descrito el modelo, pasamos a describir los experimentos realizados. Realiza una verificación dependiente del texto mediante dígitos aislados, y para una población de 10 hablantes. Compara el rendimiento de la red recurrente presentada, frente a un MLP, ambas con una misma arquitectura de tres capas, con 20 neuronas en la oculta, y 2 en la de salida. Como características extraídas de la señal de voz utilizan 10 LPCC, su derivada primera y su derivada segunda (tamaño del vector de características: 30). Estos valores de entrada a la red son normalizados entre -1 y 1, para después multiplicarles por un valor de ganancia de 50.

Cada nodo de salida de la red se hace corresponder a la clase cliente y a la clase impostor respectivamente, de forma que cuando la muestra de entrenamiento pertenece al hablante cliente, la salida deseada se corresponderá con una secuencia exponencial creciente, pasando a ser ésta decreciente en el caso de que la muestra de entrenamiento pertenezca a un impostor. Se entrena una red distinta para cada dígito y cada hablante, para lo que utiliza 6 muestras distintas del cliente y 6 de tres hablantes impostores distintos. En el proceso de aprendizaje se van alternando las muestras del cliente y de impostores, por los que cada muestra del primero se repite tres veces en cada época de entrenamiento. En prueba, la salida para una muestra de voz de origen desconocido será la media de las salidas obtenidas en el nodo perteneciente al hablante cliente. La decisión final se realiza combinando salidas de ese tipo obtenidas para conjuntos de 10 dígitos.

Utilizando como criterio de finalización del proceso de aprendizaje un número fijo de 40 épocas, la tasas de equierror obtenidas son del 7.5% para la red recurrente y del 6% para el MLP. Se puede observar que el uso de la recurrencia no mejora en rendimiento del sistema.

Otra referencia de uso de redes recurrentes las podemos encontrar en el trabajo Shrimpton y Watson (1992).

2.4.1.3.- Modelos híbridos HMM-ANN

Son técnicas que intentar aunar la capacidad de los HMMs para modelar secuencias temporales, mediante una estructura multiestado, con la capacidad tanto discriminante, como de síntesis de funciones de las ANNs. Han sido aplicadas en el caso de clasificación basada en modelos predictivos, como una extensión natural de los modelos de un solo estado, ya referenciados anteriormente. Como indicábamos allí, el problema de los modelos de un solo estado es que están suponiendo una naturaleza estacionaria en el fenómeno a estudiar, que no es cierta. La solución planteada es utilizar modeloa con M estados, donde cada estado es un predictor (MLP) diferente, especializado en una determinada parte de la muestra de voz del locutor. Este sistema puede ser considerado como un híbrido entre HMMs y MLPs. En este caso, dado un modelo, la probabilidad de que genere la secuencia de vectores de características (x_1, é , x_L) vendrá dada por:

$$P_i(x_1^L / s_1^L) = \prod_t P_i(x_t / x_T, s_t)$$

Donde s_1^L es la secuencia óptima de estados (error de predicción menor), x_T es el contexto de predicción, o sea, x_{t-1} y x_{t-2}. En entrenamiento se optimizan tanto los parámetros de las redes, como la segmentación, es decir, $P_i(x_1^L / s_1^L)$. Han sido probados tanto modelos izquierda-derecha, como ergódicos. Con estos últimos, se ha logrado alcanzar un error de identificación del 0%, con un modelo de 4 estados. Los experimentos se realizaron con 24 hablantes de la base de datos TIMIT, usando 14 segundos de voz para entrenamiento y 3 para prueba. De pruebas comparativas realizadas, el sistema presentado mejora los resultados obtenidos con modelos como VQ, HMM y ANNs discriminantes.

Un sistema similar es el presentado por Hassanein, Deng y Elmasry (1994), al que denominan modelo oculto de Markov predictivo neuronal (Neural predictive HMM). Éste consiste en un modelo izquierda-derecha donde, al igual que antes, cada estado es sustituido por un MLP que funciona como predictor para un segmento de la señal de comportamiento coasiestacionario. El entrenamiento es un proceso iterativo de dos pasos: segmentación y estimación. La segmentación consiste en la asignación de porciones de señal a cada MLP, y se realiza mediante un procedimiento de programación dinámica normal. Una vez segmentada la señal, se aplica el algoritmo de retropropagación del error para ajustar los pesos de las redes asociadas a cada segmento de voz. Este proceso iterativo se repite hasta lograr la convergencia, punto que se alcanza si el resultado de la segmentación se estabiliza o si el error de predicción cae por debajo de un determinado valor umbral. Los autores prueban con un modelo de 4 estados. Las redes neuronales tienen una arquitectura de 3 capas, con 7 neuronas en la capa de entrada, 5 en la oculta y 7 en la de salida. La representación de cada ventana de señal de voz se realiza mediante un vector de características compuesto por 7 MFCC. De la arquitectura de la red se deduce, entonces, que la ñhistoriaò utilizada en la predicción se reduce a la ventana anterior. El sistema se prueba mediante una base de datos cuyo contenido son 6 sílabas distintas, del tipo consonante-vocal, repetidas cada una 22 veces (8 para entrenamiento y 14 para prueba), por 11 hablantes diferentes (6 hombres y 5 mujeres); las muestras de voz se obtuvieron mediante un micrófono en una única sesión. En identificación dependiente del texto, siendo el ñtextoò la sílaba /di/, se obtuvo un error del 0.7%, para una población de 10 hablantes (5 h. y 5 m.), entrenando un modelo para cada hablante. En verificación dependiente del texto, siendo el ñtextoò también sílabas, pero ahora todas: se entrenó un sistema por hablante y sílaba, se obtuvieron unos resultados de 0% en falsas aceptaciones (el valor del umbral se fijo para obtener este valor) y 1.6% en falsos rechazo; ahora la población de hablantes se redujo a 3 (todos hombres), utilizando las muestras de los otros 3 hombres de la base de datos para las pruebas sobre impostores.

Un enfoque diferente lo encontramos en las denominadas ñalpha-netsò (Bridle, 1990). La técnica propuesta aplica las habilidades discriminantes del MLP a los HMM. Concretamente, una vez que el HMM está entrenado, se realiza un ajuste de las medias y las desviaciones de las gaussianas definidas en cada estado, aplicando un algoritmo de gradiente descendiente en derivadas parciales basado en el de retropropagación de los MLP. Carey, Parris y Bridle (1991) aplican esta técnica a la verificación del locutor dependiente del texto, mediante palabras clave. Por cada hablante se entrenan dos modelos, uno con muestras del hablante cliente, y otro con muestras de hablantes distintos a este. En prueba se resta la salida de ambos, si el resultado es mayor que el umbral el hablante es aceptado, siendo rechazado en caso contrario. Los experimentos realizados indican una mejora global del rendimiento del sistema tras el ñajuste discriminanteò de los parámetros: empeora la eficacia en falsos rechazos, pero en menor medida de los que mejora en falsas aceptaciones. El sistema fue probado en condiciones reales de uso: intento de acceso telefónico, mediante una secuencia de 5 digitos. Para cada hablante y dígito se crea un modelo como el descrito, de forma que es probado con 50 secuencias de 5 dígitos del propio hablante y 10 de cada uno de los otros. Cada dígito es valorado por separado, de forma que el criterio final de decisión es ñel más votadoò, o sea, si los resultados para tres o más dígitos dicen que las muestras de voz correspondientes pertenecen al hablante cliente, éste es aceptado. De las 600 pruebas realizadas, el sistema solo erró en 1.

Zeek (1996) presenta en su trabajo una técnica que podríamos denominar a caballo entre un sistema híbrido y un sistema modular. Como sistema clasificador principal utiliza un HMM, de forma que la

salida de éste: P(X/λ) (con X la muestra de voz y λ el modelo perteneciente a un determinado hablante), probabilidad no necesariamente discriminante, es procesada posteriormente por un MLP para convertirla en probabilidad a posteriori: P(λ/X). La inclusión del postprocesamiento indicado mejora notablemente el rendimiento del sistema. Así, por ejemplo, en verificación dependiente de texto, para una población de 32 hablantes, creando un modelo para cada palabra de la frase clave, se pasa de un error del 19.5% usando sólo HMMs, a un error del 6.5% con la técnica propuesta. Un problema interesante que se plantea es el de la normalización de la entrada a la red, o sea, normalizar el vector de salidas de los HMMs. El autor indica como la más eficaz de las probadas la denominada normalización estadística: el vector de entrada a la red se normaliza de forma que tenga media 0 y varianza unidad:

$$a_{ij}^{n} = \frac{a_{ij} - \mu_j}{\sigma_j}$$

Donde el subíndice i hace referencia a cada neurona de entrada a la red y el j al vector de entrada que se está normalizando. El superíndice n indica el valor normalizado.

2.4.2.- Redes neuronales en la etapa de clasificación: sistemas modulares

Uno de los primeros intentos de ñmodularizarò la etapa de decisión lo podemos encontrar en el trabajo de Rudasi y Zahorian (1991). Estos autores plantean una técnica a la que denominan ñbinary-pair neural networksò, como alternativa a los sistemas de identificación del locutor basados en una red neuronal única. El problema de este último tipo de sistemas se plantea cuando la población de hablantes se incrementa: cada vez que se añade un nuevo usuario al sistema, ha de reentrenarse la red, implicando una continuo aumento en la complejidad de ésta, lo que afecta a su rendimiento. La alternativa propuesta consiste en dividir los N hablantes a identificar en grupos de 2 (N(N-1)/2 grupos), entrenando una red distinta para cada pareja. Estas redes son del tipo MLP, con una única capa culta de 6 neuronas, y dos salidas, cada una asociada a un hablante de la pareja. Una vez entrenadas, para la asignación de una muestra de voz de origen desconocido, se plantean dos métodos distintos:

* ñGlobal soft decision searchò: la muestra de voz, o mejor, los vectores de características de ésta extraídos, sirven de entrada a todas las redes. Los valores de salida asignados a cada hablante son sumados, de forma que la asignación se realizará al locutor con el valor de la suma mayor.
* ñBinary tree searchò: las N clases (hablantes) son pareados (N/2 pares), de cada pareja se elimina al hablante cuya salida de la red asociada sea menor, de forma que a la siguiente ñrondaò solo pasan la mitad de los N locutores. Esta operación se va repitiendo hasta que sólo quede uno.

El segundo método no solo es más rápido, sino que, además, proporciona mejores resultados. El sistema se probó en comparación con la técnica de red única. Como características extraídas de la señal de voz se usaron 15 coeficientes cepstrales. El criterio de convergencia para ambas técnicas es el mismo: sobrepasar un porcentaje de aciertos umbral en el conjunto de datos de entrenamiento (este umbral varió del 30% al 50% en el caso de red única y del 65% al 75% en el enfoque de partición binaria, dependiendo del número de nodos en la capa oculta y de N). En las pruebas realizadas para identificación independiente del texto, sobre la base de datos DARPA-TIMIT, usando 5 sentencias diferentes para entrenar la red, con ambas técnicas se logró un 100% de aciertos, para una muestra de voz suficiente larga y los parámetros de las redes optimizados. La ventaja del método propuesto la encontramos cuando hay que actualizar el sistema ante la incorporación de nuevos hablantes: el tiempo necesario para entrenar la red única es superior.

En contraposición a la ventaja que acabamos de exponer, el método propuesto por Rudasi y Zahorian supone un alto coste en cuanto a número de redes, éste es proporcional al cuadrado del número de hablantes. La propuesta de Bennani (1992) pretende aunar las ventajas de la división del trabajo de clasificación entre ñexpertosò, implícito en el trabajo anterior, con un coste no tan alto en recursos. La idea básica es simple: en vez de crear parejas de hablantes, hagamos grupos más numerosos de acuerdo a algún criterio de similitud entre ellos, entrenando una red para cada uno de estos grupos, a los que los autores se refieren como ñtipologíasò.

Como características de la señal de voz se usan vectores de LPCCs, agrupando estos mediante la técnica de los k-medios. Cada hablante se asigna a la agrupación donde se encuentren la mayoría de sus vectores de características, formando así las distintas tipologías. Con los datos así obtenidos, se entrenan dos tipos de redes: una para cada tipología (la llamaremos Red de Tipología RT), encargada de diferenciar entre los distintos hablantes que la componen, y otra encargada de identificar la tipología a la que pertenece una determinada muestra de voz (la llamaremos Red Identificadora de Tipología RIT). El tipo de red neuronal empleada es el TDNN, con tres capas. Para la identificación de la muestra de voz de origen desconocido se prueban dos métodos distintos:

* ñTodo para el ganadorò. En este caso la RIT actúa a modo de puerta, o conmutador entre las distintas RTs: sólo la salida de la red correspondiente a la tipología detectada es considerada, siendo ésta la salida final del sistema.
* ñSalida ponderadaò. Para cada ventana de voz, los valores de salida de cada RT, es multiplicada por la salida correspondiente de la RIT (el mismo para todas las salidas de una determinada RT). Tras sumar los valores así obtenidos para todas las ventanas de la muestra de voz, ésta es asignada al hablante asociada a la salida con un resultado de la suma mayor.

Utilizando este segundo método para la identificación, los autores comparan el rendimiento del sistema modular propuesto, con el basado en MARMs (Multi-variate Auto-Regressive Models) (Artieres y otros, 1991). De las pruebas realizadas sobre la misma base de datos (los 2 primeros dialectos de la TIMIT, 102 hablantes), los resultados referenciados son del 100% en el sistema modular y del 95.6% en el basado en MARMs. Los autores concluyen que el método propuesto es discriminante y rápido en la fase de identificación. Además, para ésta solo se requieren muestras de voz de corta duración: 0.75s. Por contra, el tiempo de entrenamiento es muy elevado.

Haciendo referencia a la división de tareas indicada en la introducción del apartado al hablar de sistemas modulares: clasificación e integración, en los sistemas presentados las redes forman parte exclusivamente de la tarea de clasificación, realizándose la integración de los diversos resultados de formas diferentes, pero nunca con la intervención de redes neuronales. Sharma, Vermeulen y Hermansky (1998) presentan en su trabajo la situación inversa: la red neuronal, concretamente un MLP, se ocupa únicamente de la integración de las respuestas proporcionadas por dos técnicas de clasificación distintas, frente a una misma muestra de voz. La ventaja del uso de redes neuronales en esa tarea es que permite una integración no lineal de las distintas salidas. La tarea abordada es la verificación del locutor independiente de texto. Prueba el rendimiento del sistema modular en tres situaciones distintas:

* Cuando los sistemas cuyas salidas integramos tienen por separado un rendimiento similar: mezclar ambos mejora los resultados.
* Cuando uno de los dos sistemas mejora al otro: el rendimiento del sistema propuesto es similar al del mejor de ambos.
* Cada sistema a integrar funciona mejor que el otro bajo determinadas condiciones de operación. En este caso a la red que realiza la integración se le añaden nuevas entradas para las ñcondiciones de operaciónò. El sistema combinado mejora el rendimiento de ambos por separado, de forma que su eficacia se aproxima, sean cual sean las condiciones de operación, a la del mejor en cada caso.

Un enfoque distinto a los presentados, donde se diferencian claramente clasificación e integración, es el presentado más recientemente (1997), por Hadjitodorov, Boyanov y Dalakchieva, para la tarea de identificación independiente del texto. En su trabajo presentan un sistema clasificador con dos niveles, basados ambos en redes neuronales: SOMs en el primero y MLPs para el segundo, de forma que la entrada del segundo nivel es obtenida a partir de las salidas del primero. Intentan aunar la capacidad de estimación de las funciones de densidad de probabilidad de los vectores de entrada, incluso con señal ruidosa, del SOM, con la capacidad discriminante del MLP. El proceso de entrenamiento nos permite tener una visión clara del funcionamiento y características del sistema. De las muestras de voz para entrenamiento correspondientes a cada hablante se extraen las secuencias de vectores de características, 15 LPCC por ventana en este caso, que servirán de entrada al SOM. Una vez entrenado éste, se le presentan nuevamente, pero ahora con los pesos fijos, las L muestras de voz de entrenamiento de cada

hablante, evaluando, para cada hablante y muestra de entrenamiento, la frecuencia de activación de cada una de las neuronas de la red: f_{ij} , siendo i y j las coordenadas de la neurona. Se obtiene, de esta forma, para cada hablante y muestra de entrenamiento, una matriz de valores, a la que se denomina PDM (Prototype Distributing Map). Con el conjunto de PDMs obtenidos, se entrena, en el segundo nivel, un MLP por hablante. Cada una de estas redes se entrena para diferenciar a un hablante del resto, por lo que la salida deseada se fija a 1 si el PDM pertenece al hablante, y a 0 si pertenece a alguno de los restantes.

En el procedimiento descrito aparecen dos problemas. El primero es el escaso número de vectores que se tiene para entrenar el segundo nivel. Se aumenta dividiendo el conjunto de entrenamiento en T grupos disjuntos distintos, y entrenando para cada uno de estos un SOM diferente. De esta manera el número de PDMs por hablante será, ahora, de TxL El segundo problema es la posible influencia de las frecuencias menos significativas. Para evitarlo los valores de cada PDM son filtrados de la siguiente manera:

$$\text{si } 0 \le f_{ij} \le Kf_{max} \quad \text{entonces } f_{ij} = 0$$
$$\text{si } Kf_{max} \le f_{ij} \le f_{max} \quad \text{entonces } f_{ij} = f_{ij}$$

Donde $0 < K < 1$ es un coeficiente de filtrado ajustado experimentalmente, y f_{max} es el valor máximo de f_{ij}.

En prueba, de la muestra de voz de origen desconocido se obtienen T PDMs, cada uno de los cuales será procesado por cada una de las redes neuronales del segundo nivel, sumàndose cada una de las salidas así obtenidas. La muestra de voz se asignará al hablante cuya red tenga la mayor salida acumulada. Se realizó un estudio comparativo del rendimiento del sistema presentado, frente al obtenido por sistemas basados en SOMs sólo: SOM+LVQ3 con una red única de 15x15 neuronas para todos los hablantes, basados en MLPs sólo: una red por hablante con dos capas ocultas de 64 y 4 neuronas respectivamente y 1 en la de salida, y frente a sistemas basados en modelos autoregresivos (AR-models). Las condiciones experimentales, las mismas para todos los sistemas, son:

* Pruebas con señal limpia: 68 hablante (35 hombres y 33 mujeres), 2 sentencias de 4 a 7 segundos por hablante para entrenamiento, también dos sentencias de prueba por hablante, distintas a las de entrenamiento y obtenidas en sesiones distintas a las de éste.
* Pruebas con señal telefónica: 92 hablantes (48 hombre y 44 mujeres), 3 sentencias de 4 a 7 segundos por hablante para entramiento, también 3 sentencias por hablante para prueba, con las mismas características que las del punto anterior.

Con señal limpia los errores de identificación fueron del 3.67% con MLP solo, del 2.94% con SOM+LVQ solo, del 3.67% con los modelos autoregresivos y del 2.2% con el sistema en dos niveles propuesto. Para la señal telefónica los errores obtenidos, puestos en el mismo orden que antes, fueron: 6.15%, 5.43%, 5.79% y 2.17%.

2.4.3.- Otras aplicaciones

Las redes neuronales también han sido utilizadas en la etapa de extracción de características, previa a la de clasificación. En general, se busca realizar un tratamiento de los vectores de características de forma que se potencie la influencia del hablante en ellos. Un ejemplo de esto lo podemos encontrar en el trabajo de Konig y otros (1998), donde, mediante un MLP se busca transformar el espacio de características, de manera que se disminuya la dimensionalidad, al tiempo que se incremente la separación entre los vectores pertenecientes a las distintas clases. Realizan lo que denominan un análisis discriminante no lineal (NLDA) de los datos, para lo cual entrenan una red MLP con 5 capas: 500 neuronas en la primera capa oculta, 34 en la segunda y 500 en la tercera; la capa de salida tiene tantas neuronas como hablantes a reconocer, de forma que, en entrenamiento, solo la salida correspondiente al hablante al que pertenece el vector de entrada será la que se active. Una vez concluido el entrenamiento, se eliminan las dos ultimas capas de la red, de forma que la salida de la segunda capa oculta (con 34 neuronas) será la utilizada como nuevo vector de características (transformación no lineal del original) usado para el reconocimiento del locutor: será la entrada a la etapa de clasificación, realizada en este caso mediante mezcla de gaussianas.

El entrenamiento de la red que realiza el NLDA se realiza mediante un conjunto de muestras de entrenamiento que no tiene porque ser el mismo que el que se usará posteriormente para entrenar los modelos de mezcla de gaussianas. De las pruebas realizadas, usando el corpus de evaluación para el reconocimiento del locutor NIST 1997, se deduce que el sistema propuesto no mejora al clásico basado en el uso de coeficientes cepstrales directamente (sin transformar) en la etapa de clasificación. Sin embargo una combinación lineal de ambos, con pesos 0.3 para la salida del basado en NLDA y 0.7 para el más clásico, si que mejora el produce una mejora en el reconocimiento.

Otro ejemplo de uso de redes en el acondicionamiento de las características extraídas de la señal de audio, previo a su clasificación, lo podemos encontrar en el trabajo de Lin y otros (1994). En este caso utilizan un MLP con 3 capas (12 neuronas en las capas de entrada y salida, y 8 en la oculta) como filtro, para intentar eliminar, en lo posible, los errores debidos a las diferentes condiciones en que pueden ser obtenidas las muestras de voz de entrenamiento y prueba, concretamente, la red trata de adaptar las condiciones de prueba a las de entrenamiento. Los resultados muestran una notable mejora en los resultados con la introducción de la etapa de filtrado, cuando las condiciones de entrenamiento y prueba son distintas.

BIBLIOGRAFÍA

Anderson T. R. y Patterson R. D. (1994). ñSpeaker Recognition with Auditory Image Model and Self Organizing Feature Maps: A Comparison With Traditional Techniquesò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 153-156, Martigny (Suiza), 1994.

Artieres T., Bennani Y., Gallinari P. y Montacie C. (1991). ñConnectionist and Conventional Models for Free Text Talker Identificationò. Neuro-Nimes, Francia, 1991.

Artieres T. y Gallinari (1993). ñ Neural Models for Extracting Speaker Characteristics in Speech Modelization Systemsò. Proc. EUROSPEECH, Berlin (Alemania), 1993.

Bennani Y., Fogelman F. y Gallinary P. (1990). ñA Connectionist Approach for Automatic Speaker Identificationò. Proc. IEEE ICASSP, S5.2, pp. 265-268, Albuquerque, New Mexico (USA), 1990.

Bennani Y. (1992). ñ Speaker Identification Through a Modular Connectionist Architecture: Evaluation on the TIMIT Databaseò. Proc. ICLP, S4.4, pp. 607-610, Banff (Canada), 1992.

Bimbot F., Mathan L., De Lima A. y Chollet G. (1992). ñStandard and Target Driven AR-Vector Models for Speech Analysis and Speaker Recognitionò. Proc. IEEE ICASSP, Vol. 2, pp. 5-8, San Francisco (USA), 1992.

Boves L. (1998). ñCommercial Applications of Speaker Verification: Overview and Critical Success Factorsò. Proc. RLA2C Workshop on Speaker Recognition and Its Commercial and Forensic Applications, pp.150- 159, Avignon (Francia), 1998.

Bridle J. (1990) ñAlpha-Net: a Recurrent Neural Network Architecture with a Hidden Markov Modelò. Speech Communication, special Neurospeech issue, 1990.

Carey M. J., Parris E. S. y Bridle J. S. (1991). ñA Speaker Verification System Using Alpha-Netsò. Proc. IEEE ICASSP, S6.8, pp. 397-400, Toronto (Canada), 1991.

Fredrickson S. E. y Tarassenko L. (1994). ñRadial Basis Functions for Speaker Identificationò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 107-110, Martigny (Suiza), 1994.

Furui S., Itakura F. y Saito S. (1972). ñTalker Recognition by Longtime Averaged Speech Spectrumò. Trans. IECE, 55-A, Vol. 1, No. 10, pp. 549-556, 1981.

Furui S. (1981). ñCepstral Analysis Technique for Automatic Speaker Verificationò. IEEE Trans. Acoust. Speech Signal Processing, Vol. 29, No. 2, pp. 254-272, 1981.

Gong Y. y Haton J.-P. (1994). ñNon-Linear Interpolation Methods for Speaker Recognition and Verificationò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 23-26, Martigny (Suiza), 1994.

Hadjitorov S., Boyanov B. y Dalakchieva N. (1997). ñA Two Level Classifier for Text Independent Speaker Identificationò. Speech Communication, Vol 21, No. 3, pp. 209-217, Abril 1997.

Hassanain K., Deng L. y Elmasry M. I. (1994). ñA Neural Predictive Hidden Markov Model for Speaker Recognitionò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 115-118, Martigny (Suiza), 1994.

Hattori H. (1992). ñText Independent Speaker Recognition Using Neural Networksò. Proc. IEEE ICASSP, Vol. 2, pp. 153-156, San Francisco (USA), 1992.
Hattori H. (1994). ñText Independent Speaker Verification Using Neural Networksò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 103-106, Martigny (Suiza), 1994.
Hennebert J. y Delacretaz D. P. (1998). ñPhoneme Based Text Prompted Speaker Verification with Multi Layer Perceptronsò. Proc. RLA2C Workshop on Speaker Recognition and Its Commercial and Forensic Applications, pp.55-58, Avignon (Francia), 1998.
Hermansky H. y Narendranath M. (1998). ñSpeaker Verification Using Speaker-Specific Mappingsò. Proc. RLA2C Workshop on Speaker Recognition and Its Commercial and Forensic Applications, pp.111-114, Avignon (Francia), 1998.
Konig Y., Heck L., Weintraub M. y Sonmez K. (1998). ñNonlinear Discriminant Feature Extraction for Robust Text Independent Speaker Recognitionò. Proc. RLA2C Workshop on Speaker Recognition and Its Commercial and Forensic Applications, pp.72-75, Avignon (Francia), 1998.
Labanova M. A. y Raev A. N. (1998). ñSpeaker Verification Accounting the Formant Behaviour and Phonetic Representation of Enrolled Speechò. Proc. RLA2C Workshop on Speaker Recognition and Its Commercial and Forensic Applications, pp.37-39, Avignon (Francia), 1998.
Lastrucci L., Gori M. y Soda G. (1994). ñNeural Autoassociators for Phoneme-Based Speaker Verificationò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 189-192, Martigny (Suiza), 1994.
Levin E. (1990). ñModelling Time Varying Systems Using Hidden Control Neural Architetureò. NIPS 3, pp. 147-154, 1990.
Li K.-P. y Wrench E. H. (1983). ñ An approach to text Independent Speaker Recognition with Short Utterancesò. Proc. IEEE ICASP, 12.9, pp. 555-558, 1983.
Lin Q., Jan E., Che C. y Flanagan J. (1994). ñSpeaker Identification in Teleconferencing Environments Using Microphone Arrays and Neural Networksò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 235-238, Martigny (Suiza), 1994.
Markel J. D., Oshika B. T. y Gray A. H. (1977). ñLong-Term Feature Averaging for Speaker Recognitionò. Proc. IEEE Trans. Acoust. Speech Signal Processing, Vol. ASSP-27, No.1, pp. 330-337, 1977.
Matsui T., Furui S. (1992). ñComparison of text Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMsò. Proc. IEEE ICASSP, Vol. 2, pp. 157-160, San Francisco (USA), 1992.
Montacié C. y Le Floch J.-L. (1992). ñAR-Vector Models for Free Text Speaker Recognitionò. ICSLP, 1992.
Narendranath M., Murthy H. A., Rajendran S. y Yegnanarayana B. (1994). ñVoice Conversion Using Artificial Neural Networksò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 197-200, Martigny (Suiza), 1994.
Oglesby J. y Mason J. S. (1990). ñOptimization of Neural Models for Speaker Identificationò. Proc. IEEE ICASSP, S5-1, pp. 261-264, Albuquerque, New Mexico (USA), 1990.
Oglesby J. y Mason J. S. (1991). ñRadial Basis Function Networks for Speaker Recognitionò. Proc. IEEE ICASSP, S6.7, pp. 393-396, Toronto (Canada), 1991.
Portiz A. B. (1982). ñLinear Prediction Hidden Markov Models and the Speech Signalò. Proc. IEEE ICASSP, S11.5, pp. 1291-1294, 1990.
Rose R. y Reynolds R. A. (1990). ñText Independent Speaker Identification Using Automatic Acoustic Segmentationò. Proc. IEEE ICASSP, S51.10, pp. 293-296, Albuquerque, New Mexico (USA), 1990.
Rudasi L. y Zahorian S. A. (1991). ñText Independent Talker Identification with Neural Networksò. Proc. IEEE ICASP, S6.6, pp. 389-392, Toronto (Canada), 1991.
Sharma S., Vermeulen P. y Hermansky H. (1998). ñCombining Information from Multiple Classifiers for Speaker Verificationò. Proc. RLA2C Workshop on Speaker Recognition and Its Commercial and Forensic Applications, pp.115-118, Avignon (Francia), 1998.
Shikano K. (1985). ñText Independent Speaker Recognition Experiments Using Codebooks in Vector Quantizationò. J. Acoust. Soc. Am. (abstract), Suppl. 1, No. 77, p. S11, 1985.
Shrimpton D. y Watson B. (1992). ñComparison of Recurrent Neural Networks Architectures for Speaker Identity Verificationò. Proc. SST, pp. 460-464, 1992.

Silva H., Vivaracho C. E., Alonso L. and Cardeñoso V. (1998), ñSpeaker Verification: A Comparison Between ANNs and HMMs approachò, Proc. Workshop on Arificial Neural Networks: Current Trends and Applications. 4th World Congress on Expert Systems, Ciudad de Mexico (Mexico), 1998.

Sorensen H. B. D. y Hartman U. (1993). ñPi-Sigma and Hidden Control Based Self Structuring Models for Text Independent Speaker Recognitionò. Proc. IEEE ICASSP, pp. 537-540, Minneapolis (USA), 1993.

Tsoi A. C., Shrimpton D., Watson B. y Back A. (1994). ñApplication of Neural Network Techniques to Speaker Verificationò. Proc. Workshop on Automatic Speaker Recognition Identification and Verification, pp. 143-152, Martigny (Suiza), 1994.

Vivaracho C. E., Alonso L. y Sahelices B. (1994). ñSpeaker Recongnition by Mean Energy per Frequency Band. First Approximationò, Proc. 12th IASTED International Conference on Applied Informatics, pp. 75-77, Annecy (Francia), 1994.

Zeek E. J. (1996). ñSpeaker Recognition by Hidden Markov Modelsò. Tesis presentada en la escuela de ingenieria del instituto de tecnología del ejército del aire (USA), 1996.

Zheng Y.-C. y Yuan B.-Z. (1988). ñText Dependent Speaker Identification Using Circular Hidden Markov Modelsò. Proc. ICASSP, S13.3, pp. 580-582, 1988.