Título: Reconocimiento automático del habla con redes neuronales artificiales

Autor: Diego H. Milone

Resumen: En la última década se han investigado intensamente los fundamentos y aplicaciones de las redes neuronales artificiales. Simultanea y muchas veces concurrentemente, se ha invertido un gran esfuerzo en el área del reconocimiento automático del habla. Este trabajo presenta una extensa revisión y un análisis crítico de la aplicación de redes neuronales artificiales al reconocimiento automático del habla. Se mencionan los antecedentes de la década de los 80 y se describen los trabajos que constituyen una aplicación directa de técnicas clásicas de redes neuronales artificiales. Se discuten aquellos trabajos en los que se han desarrollado nuevas arquitecturas neuronales, orientadas a resolver el problema del reconocimiento automático del habla. Dado que los modelos ocultos de Markov han tenido el mayor éxito en esta tarea, se presentan también numerosos trabajos que consisten en la combinación de redes neuronales artificiales y modelos ocultos de Markov. Finalmente se realiza un análisis comparativo de los diferentes enfoques y los puntos clave que guían las investigaciones entorno a nuevos paradigmas para resolver este problema aún abierto.

Palabras clave: redes neuronales artificiales, reconocimiento automático del habla, híbridos de redes neuronales artificiales y modelos ocultos de Markov.

I. Introducción

Este artículo contiene el resultado de una recopilación y análisis crítico de material bibliográfico relacionado con el reconocimiento automático del habla (ASR[1]) mediante redes neuronales artificiales (ANN). La recopilación se ha realizado en diferentes publicaciones científicas tanto en el área de redes neuronales como en el área del procesamiento del habla y del lenguaje.

I.1. Antecedentes

Para comenzar es necesario realizar un breve comentario acerca de algunos trabajos pioneros en ANN-ASR durante la década del 80. Dentro de éstos se incluyen los aportes de:

- *Teuvo Kohonen*, en los mapas autoorganizativos (SOM) (**Kohonen** y cols., 1984) (**Kohonen**, 1990). Estos trabajos se desarrollaron en el Laboratory of Computer and Information Science de la Helsinki University of Tecnology (Finlandia). Una revisión muy completa de los aportes de Teuvo Kohonen se puede encontrar en su libro (**Kohonen**, 1995).
- *Richard Lippman*, en perceptrones multicapa (MLP) y combinaciones de ANN y modelos ocultos de Markov (HMM) (**Lippmann**, 1987). Estos trabajos fueron desarrollados en el MIT Lincoln Laboratory (EEUU).
- *Hervé Bourlard*, principalmente relacionados con MLP (**Bourlard** y **Wellekens**, 1989) e híbridos entre MLP y HMM (una revisión más completa se puede ver en **Bourlard** y **Wellekens**, 1990). Estos trabajos fueron desarrollados en la Faculté Polytechnique de Mons (Bélgica).
- *Alex Weibel*, en las redes neuronales con retardos en el tiempo (TDNN) (**Waibel** y cols., 1989a) (**Waibel** y cols., 1989b). Estos trabajos se desarrollaron en el Computer Science Department de la Carnegie Mellon University (EEUU).

Evidentemente estos no son los únicos autores relevantes de la década del 80. Hay que considerar también a muchos de sus coautores y otros trabajos como por ejemplo los de **Elman** (1988) y **Jordan** (1986), en las redes recurrentes que luego llevaron sus nombres.

Cabe aclarar en este punto que se ha evitado sistemáticamente caer en la descripción de los detalles de implementación y formalismos de cada artículo. Por un lado, porque para eso están los artículos en si mismos y por otro, porque la extensión que este trabajo alcanzaría lo haría probablemente ilegible y seguramente aburrido.

I.2. Criterios de selección

Para la selección se han considerado –casi exclusivamente– artículos de revistas debido a que los artículos en congresos y conferencias generalmente no reportan la suficiente información para el análisis y contienen sólo resultados preliminares.

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

Debido a la gran cantidad de material encontrado (más de 120 artículos inicialmente) y a la extensión sugerida para este trabajo, se hizo una selección de aquellos trabajos que realizan un aporte más significativo. La primera selección se basó en el resumen y las conclusiones del trabajo. La segunda selección fue posterior a la lectura del artículo y la tercera después de un análisis en profundidad de cada uno. Finalmente han sobrevivido algo más de 60 artículos aunque en las referencias también se incluyen otros artículos relacionados con antecedentes, revisiones, tutoriales y bases de datos.

I.3. Criterios de clasificación

No es simple realizar una clasificación de tantos trabajos sin encontrarse muchas veces en situaciones comprometidas. La clasificación elegida, que determina la estructura de este artículo, sigue la idea de que tratamos con una aplicación de ANN. Por esta razón, en la clasificación se ha puesto especial atención en las arquitecturas neuronales para dejar en segundo plano los aspectos relacionados con ASR. De esta forma, las 3 grandes secciones descriptivas tratan sobre: la aplicación de modelos neuronales conocidos al ASR, nuevos modelos neuronales diseñados para ASR y modelos que combinan ANN y HMM.

En cuanto a lo relacionado con el ASR se ha tratado de seguir un esquema único que contemple los aspectos más importantes y permita un análisis comparativo. Para esto se han tenido en cuenta aspectos como: el nivel al que se realiza el reconocimiento, las características de los hablantes y la cobertura del problema de ASR que se alcanza. El nivel al que se realiza el reconocimiento distingue más que nada entre reconocimiento de fonemas o palabras, aisladas o en voz continua. En el caso de los hablantes interesa si el reconocimiento es dependiente del locutor (con locutor único o múltiples locutores) o independiente del locutor, y cuantos locutores se utilizaron en las pruebas. Finalmente la cobertura del problema del problema de ASR es una valoración global con la que se intenta brindar una idea de cuantos de los aspectos importantes del ASR han sido abordados satisfactoriamente por los autores[2].

Para los artículos más interesantes se provee una tabla con un resumen que incluye estos aspectos del ASR y las características estructurales más importantes del sistema propuesto[3]: el modelo, las entradas y su preprocesamiento, las salidas o método de clasificación y el método de entrenamiento. Para terminar, se consideran en cada tabla datos metodológicos como: las condiciones de los experimentos realizados, las bases de datos utilizadas y una valoración de los resultados. Esta valoración no se trata simplemente de una medida lingüística de los resultados porcentuales, sino que puede verse disminuida en casos en que no estén claros o se consideren insuficientes los métodos de validación utilizados.

I.4. Bases de datos de dominio público

Un aspecto que permite valorar con más certeza los resultados de un artículo es la base de datos que se utilizó. Muchos autores construyen sus propias bases de datos (por razones que no interesa analizar aquí) y esto hace que sea difícil saber, a pesar de las descripciones de ellos mismos, qué tan compleja es la tarea de reconocimiento en esa base de datos. Por esta razón es difícil valorar los resultados y casi imposible reproducirlos. En el otro extremos están los artículos que realizan las pruebas con bases de datos de dominio público (gratuitas o comerciales). En este caso se conoce muy bien la complejidad de la tarea de reconocimiento e incluso se posee acceso a la misma base de datos para realizar experimentos comparativos. Las bases de datos que se incluyen en los artículos analizados son: TIMIT (**Zue** y cols., 1990), PB (**Peterson** y **Barney**, 1952), DARPA-1000 (**Price** y cols., 1988), PhoneBook (**Pitrelli**, 1995) y CSLU-OGI_30K (**Cole** y cols., 1995). No haremos aquí una descripción amplia sino sólo una breve referencia a cada una. TIMIT es una base de datos de voz continua, con más de 600 hablantes (existen varias versiones) y en gran parte segmentada en fonemas. La base PB es una antigua base de datos que ha sido utilizada para pruebas de clasificación y solamente contiene las formantes de 10 vocales para 76 hablantes. La base de datos DARPA-1000 contiene 1000 palabras de 109 hablantes y unas 300 frases de baja perplejidad. Finalmente el CSLU-OGI_30K posee registros telefónicos de más de 30000 frases de números aislados y conectados. En todos los casos se trata de bases de datos en inglés.

I.5. Artículos tutoriales de interés y revisiones

Para concluir con la introducción se hará referencia a diversos artículos que pueden complementar la lectura del presente.

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

Como el trabajo trata principalmente de ANN, se ha intentado no abrumar al lector con demasiados detalles relativos a HMM (cuando esto fue posible). En relación a los HMM hay una referencia clásica para introducción: el artículo de **Rabiner** y **Juang** (1986). Para un análisis más profundo se recomienda el libro de los mismos autores (**Rabiner** y **Juang**, 1993) o el libro de **Jelinek** (1999).
En relación a las redes neuronales se puede encontrar una introducción corta en los artículos de **Lippmann** (1987) y **Widrow** y **Lehr** (1990). Para mayor profundidad se pueden consultar los libros de **Freeman** y **Skapura** (1991) y **Haykin** (1994). En particular para redes neuronales recurrentes hay dos artículos de revisión y generalización: (**Tsoi** y **Back**, 1994) y (**Tsoi** y **Back**, 1997).
En cuestiones relacionadas con el procesamiento de la voz se pueden consultar los libros de **Deller** y cols. (1993) y **Rabiner** y **Gold** (1975).
Existen otras revisiones acerca de ANN-ASR y en esta sección se hará una breve mención a cada una. En primer lugar se pueden consultar varios artículos de Hervé Bourlard, que siempre contienen una amplia revisión comparativa de trabajos anteriores ((**Bourlard** y **Morgan**, 1993), (**Bourlard** y **Wellekens**, 1990) y (**Bourlard** y cols., 1992)). En estos artículos principalmente se trata la combinación ANN-HMM. En **Bourlard** y cols. (1996) en particular, se puede encontrar un trabajo muy estimulante en cuanto a la búsqueda de nuevas alternativas para solucionar el problema de ASR. También Richard Lippmann ha publicado interesantes artículos iniciales que proveen una visión global de la utilización de ANN para ASR (**Lippmann**, 1989) e interesantes comparaciones entre las capacidades humanas y las de los sistemas actuales para ASR (**Lippmann**, 1997). Finalmente, un trabajo más actual, más bien enfocado hacia las combinaciones ANN-HMM se encuentra en (**Wiliñski** y cols., 1998). Aunque es un trabajo de revisión amplio, probablemente adolece del problema de estar basado, casi por completo, en artículos de congresos y conferencias.

II. Aplicaciones de ANN para ASR

En esta sección se presentan, más brevemente que en las siguientes, trabajos con arquitecturas neuronales clásicas aplicadas al ASR. Se eligió esta como primera sección por varias razones. Una es que la complejidad de los sistemas propuestos es baja. También, en términos generales, se trata de trabajos de los primeros años de la década del 90, lo que nos permite aproximar un orden cronológico. Y por ultimo, en cuanto a los resultados y cobertura del problema de ASR, estos trabajos pueden ser considerados como un buen comienzo.
En esta sección se ha preferido subclasificar los trabajos de acuerdo a su objetivo y no a su arquitectura neuronal ya que, en cuanto a esto último, no aportan mayores novedades.

II.1. Clasificación de vocales

En el trabajo de **Knagenhjelm** y **Brauer** (1990) se compara el desempeño de dos sistemas neuronales para la clasificación de vocales. Los clasificadores son un MLP (40-20-12)[4] y un SOM (16x16). Para las pruebas se utiliza una base de datos propia con un solo locutor profesional masculino. 12 vocales del sueco son segmentadas y extraidas manualmente para generar un vector de coeficientes cepstrales (CC) cada 16 ms (**Deller** y cols., 1993). Un tercer sistema donde las salidas del SOM alimentan un perceptrón de 2 capas es evaluado pero, si bien los resultados superan al SOM, son inferiores al MLP. No está clara la significancia estadística de los resultados ya que no se analiza y la diferencia es de un punto entre cada sistema sobre un porcentaje de reconocimiento medio de 82%. Además, el hecho de que el hablante sea profesional hace que los resultados sean menos generalizables. La segmentación manual y la clasificación puramente estática hacen que el trabajo posea una aplicación muy reducida en el campo del ASR.
Irino y **Kawahara** (1990) proponen un nuevo método para entrenar un MLP para la clasificación de vocales. El método de análisis de regresión lineal (MRA) se corresponde con un perceptrón lineal. Cuando se incluyen neuronas no lineales entonces es necesario ampliar el MRA a su versión no lineal y así en el artículo se presenta como alternativa un caso particular denominado modelo logístico múltiple (MLM).
Las pruebas de reconocimiento incluyen solamente 5 vocales y están más orientadas a la comparación de los métodos que al ASR. Dada la configuración propuesta en el diseño de la segunda capa de neuronas, el modelo es difícilmente ampliable a la totalidad de fonemas. Además no deja de tratarse de un método de clasificación estática, lo que constituye una gran desventaja para tratar el problema de ASR.
Cosi y cols., (1990) describen un método de clasificación de 10 vocales del inglés basado en MLP. El sistema consiste en dos módulos principales, donde el primero es un modelo de oído con el que

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

se realiza la extracción de características. El segundo módulo es un MLP para el reconocimiento de vocales, independiente del hablante pero con contexto fijo (lo que simplifica significativamente el problema). El algoritmo de entrenamiento es el de retropropagación (BP) clásico y se obtienen buenos resultados mediante una codificación binaria de las salidas deseadas para la clasificación de las 10 vocales. Una particularidad extraña del entrenamiento es que se itera hasta lograr un 0% de error. Esto podría suponer un problema de sobreentrenamiento y la consecuente reducción del desempeño sobre el conjunto de prueba.

En otra serie de pruebas se utilizan tres MLP para realizar una clasificación en función de las características articulatorias de generación de las vocales. Uno de los MLP se encarga de la clasificación del lugar de la articulación, otro de la forma de la articulación y el último de la tensión articulatoria. Teniendo en cuenta estas tres clasificaciones parciales se llega a un reconocimiento global del 89.4%. Un criterio adicional (basado en centroides) se utiliza cuando la conjunción de las tres salidas de los MLP no se corresponde con ninguna de las vocales a clasificar. El trabajo es extenso y claro pero la cobertura del problema de ASR es muy baja.

Por último, y ya sobre el final de la década, **Zahorian** y **Nossair** (1999) proponen un método de partición del proceso de clasificación en el que la clasificación de *N* clases es subdividida en múltiples clasificaciones en 2 clases. En este trabajo también se desarrolla una nueva técnica de extracción de características para ASR (con resultado similar a los CC pero con mayor flexibilidad en el control de la resolución tiempo-frecuencia). El sistema inteligente se compone por múltiples MLP entrenados con BP. Se clasifican vocales de la base de datos TIMIT y si bien los experimentos están muy bien descriptos, los resultados son buenos y el análisis de ellos es amplio y adecuado, la cobertura del problema de ASR es muy baja por lo que no se considera oportuno presentar más detalles en este trabajo.

II.2. Clasificación de fonemas

Comenzando a ampliar la cantidad clases **Kong** y **Kosko** (1991) proponen unas nuevas ANN de aprendizaje competitivo diferencial (DCL) para la estimación de centroides en ASR. La estructura propuesta es similar a la de un SOM lineal con conexión total e inhibición lateral. Lo más novedoso de esta propuesta es el método de aprendizaje utilizado que consiste en una combinación de aprendizaje competitivo y hebbiano diferencial. Realmente la aplicación al ASR es muy simple y los resultados que se presentan no son útiles para la comparación con otros trabajos ya que se trata de 9 fonemas fácilmente clasificables (/*a*/, /*e*/, /*i*/, /*o*/, /*u*/, /*f*/, /*s*/, /*n*/, /*t*/). Además, la base de datos es registrada por los mismos autores y para un solo hablante masculino con algunos casos generados artificialmente. En cuanto a los resultados, se puede observar que las vocales son bien reconocidas pero, a pesar de no tratarse consonantes difíciles de clasificar, existe un alto porcentaje de error para la /*f*/ y la /*t*/. Evidentemente se trata de un sistema de clasificación estático que no maneja, desde ningún punto de vista, aspectos temporales de la clasificación. El preprocesamiento de la señal de voz tiene la particularidad de dividir un análisis tiempo-frecuencia estándar (transformada de Fourier de tiempo corto (STFT) con ventana de Hamming) en diferentes bandas de interés, modificando la resolución frecuencial de cada una (6 coeficientes entre 200Hz y 3 kHz y 13 entre 3 y 5kHz). Se podría decir que la cobertura del problema de ASR no supera el 10%.

En **Wu** y cols. (1992) se propone la utilización de un SOM (20x20) para la clasificación de fonemas chinos. Realmente los autores no aportan ninguna novedad a la técnica del momento sino que extienden los trabajos de Kohonen a un nuevo conjunto de fonemas. La región de influencia para la neurona ganadora es de 3x3 (burbuja de disparo). Para el preprocesamiento se utilizan 17 coeficientes espectrales (SC). Los resultados no se especifican claramente en función de porcentajes de error o alguna medida similar.

Con una visión más acertada sobre la dinámica de la voz, **Hanes** y cols. (1994) realizan un reconocimiento de fonemas mediante RNN (ver también **Wang** y cols. (1996)). La arquitectura neuronal utilizada responde al modelo sencillo de Elman. Si bien los porcentajes de resultados parecen buenos, hay que considerar muchos puntos débiles del trabajo: los fonemas reconocidos son sólo 6 (en contexto), la base de datos es registrada por ellos mismos y de un solo hablante masculino, como parámetros se utilizan las trayectorias de las tres primeras formantes (lo que explica que los resultados sean notoriamente mejores para vocales que para consonantes) y el método de normalización utilizado se basa en que la segmentación es conocida de antemano. Finalmente hay que destacar que el método de validación y obtención de los resultados, basado en utilizar la red de Elman como predictor de las trayectorias de formantes, es bastante dudoso y los

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

mismos autores reconocen sobre el final del artículo que sería bueno encontrar un nuevo método para determinar la precisión de la red.

II.3. Clasificación de palabras (aisladas)

Un MLP parcialmente conectado (PCMN) se presenta como alternativa para el reconocimiento de palabras aisladas en el trabajo de **Ye** y cols. (1990). Principalmente los autores hacen hincapié sobre la capacidad de generalización de una red PCMN dado su poder de interpolación sobre una base de datos con pocos ejemplos. La clasificación sigue siendo de tipo estática, introduciendo a la entrada el vector de características para el tiempo actual y otros de contexto. Se examinan diferentes configuraciones donde ciertos elementos de la entrada no son conectados a algunos de los elementos del vector de características o su contexto. El algoritmo de entrenamiento es BP pero, y esto es importante, no se presenta un método general para especificar las conexiones ausentes en el PCMN.

Las pruebas se realizan con 10 números en chino pronunciados por un hablante masculino y otro femenino. Para cada vector de características se realiza un preprocesamiento para obtener 10 coeficientes cepstrales en escala de mel (MFCC). Los resultados finales alcanzan un 72% para la red totalmente conectada y un 91% para la parcialmente conectada, después de probar muchas diferentes configuraciones de conexionado parcial. En una de las conclusiones los autores dicen que "esta estrategia de conexionado parcial es una buena alternativa para el manejo eficiente de la información temporal". Sin embargo, lo que se obtiene es una red con mayores capacidades de generalización pero difícilmente este clasificador estático tenga posibilidades de inferir la estructura temporal de la secuencia mucho más allá de la ventana de tiempo que obtiene como entrada. Tampoco queda claro por qué los autores deciden incorporar en las salidas del PCMN una distinción entre si el número lo dijo un hombre o una mujer cuando, originalmente, no se plantea como objetivo determinar el sexo del hablante.

También en la línea de reconocimiento de palabras aisladas, **Gramss** y **Strube** (1990) proponen el reconocimiento de números aislados en alemán mediante un sistema de preprocesamiento diseñado psicoacústicamente y un sencillo perceptrón de 730 neuronas (sin capa oculta). Las 10 salidas se corresponden una con cada número. Los porcentajes de reconocimiento son elevados pero la cobertura del problema de ASR es muy baja.

Woodland y **Smyth** (1990) comparan 4 métodos de clasificación estática para el reconocimiento de dos palabras en inglés ("yes" y "no"). Se trata de clasificadores basados en *k*-medias, mezclas de guassianas, MLP y SOM. Se analizan varias configuraciones, los resultados comparativos y el costo computacional de cada técnica. Los mejores resultados los da el MLP, con la estructura más simple pero con el mayor tiempo de entrenamiento.

Las redes recurrentes caóticas (CRNN) son aplicadas al reconocimiento de dígitos y sílabas aisladas en el trabajo de **Ryeu** y **Chung** (1996). Estas redes enfatizan las características no lineales del modelo de neurona. La arquitectura consiste en una capa de entrada que simplemente cumple con la función de distribuir los patrones y una capa recurrente con diversas realimentaciones y sus factores de peso correspondientes. Una realimentación se produce desde la salida lineal de cada neurona oculta hacia si misma. Otra se realiza desde después de la no-linealidad y también hacia la misma neurona. Y las últimas van desde el mismo lugar hacia la entrada, conformando así una extensión del vector de entradas. Las primeras retroalimentaciones hacia las mismas neuronas tienen la función de controlar su dinámica temporal de activación junto con otros parámetros de umbral. Los factores que modifican la influencia de estas realimentaciones son determinados de forma empírica. Las realimentaciones hacia la entrada forman parte del enfoque más clásico de RNN y los pesos entre todas las entradas y la capa oculta son ajustados mediante un algoritmo basado en el descenso del gradiente de error. Error que es medido y acumulado a lo largo del tiempo durante la respuesta dinámica de la CRNN. El primer experimento se realiza con dígitos aislados en coreano. La base de datos se compone de 35 ejemplos de cada dígito pronunciados por un único hablante. Los vectores de características consisten en 10 MFCC. Los resultados alcanzan el 96.3% contrastando con un 92.3% de una RNN sin la dinámica caótica.

El segundo experimento consiste en la clasificación de 672 monosilábicos del coreano. Con una sola CRNN para todas las sílabas los porcentajes de error son muy altos. Por esta razón se separa la clasificación en tres grupos. El primero contiene las sílabas cortas de tres fonemas, el segundo contiene las sílabas largas de tres fonemas y el tercero las sílabas de dos fonemas. Para la preclasificación en estos tres grupos se utiliza un algoritmo basado en reglas. En primer lugar se

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

separa el primer grupo considerando simplemente la duración de la sílaba. Para todas las sílabas de este grupo se aplica una normalización de manera que queden de la misma longitud. A todas las demás sílabas se les aplica otra normalización de tiempo y luego se separan los dos grupos según un criterio basado en máximos de energía y umbrales empíricamente fijados. Los resultados finales no superan el 90%.

Resumen para	(**Ryeu** y **Chung**, 1996)
Modelo	CRNN: RNN con énfasis en realimentaciones que generan un comportamiento caótico.
Entradas	MFCC, preclasificación en grupos de fonemas basada en reglas
Salidas	Predicción temporal.
Entrenamiento	Varios parámetros fijados empíricamente y los pesos por gradiente descendiente
Experimentos	Dígitos y monosílabos aislados
Base de datos	Propia, un hablante coreano
Reconocimiento	Palabras aisladas
Valoración de resultados	Regulares
Cobertura ASR	40%

Existen dos consideraciones finales en relación a este artículo. La primera es que realmente el manejo del tiempo y la segmentación es totalmente artificioso y ajeno a la CRNN. La segunda se resume en la pregunta: ¿por qué los autores deciden utilizar un sistema caótico para modelar una señal que no lo es? (al respecto véase el trabajo de **Banbrook** y cols. (1999)).

Kuhn y cols. (1990) realizan un reconocimiento de 9 nombres de letras en inglés por medio de una red recurrente y con retardos temporales. El problema no es tan simple como el de reconocimiento de vocales porque el nombre de varias letras es bastante fácil de confundir. La estructura neuronal posee 17 nodos en la entrada y 9 en la capa oculta. La entrada está completamente conectada hacia la capa oculta con cuatro retardos temporales. La capa oculta se conecta a la salida con 3 retardos temporales (1, 3 y 5) y los nodos de salida se conectan recurrentemente hacia si mismos con un retardo temporal. El algoritmo de entrenamiento se basa en la minimización de un criterio de error a través de gradiente descendiente.

La base de datos es propia, con hablantes masculinos y un preprocesamiento basado en SC en escala de mel (MSC) y coeficientes delta (ΔC) (derivadas en el tiempo). En los resultados se compara una versión no recurrente de la ANN con la versión recurrente que funciona algo mejor. Los mejores resultados no alcanzan a los obtenidos en circunstancias similares con HMM o TDNN.

II.4. Otras aplicaciones

En el trabajo de **Watrous** (1993) se propone un método de normalización y adaptación al hablante basada en ANN. Antes que nada hay que notar que las pruebas se realizaron solamente con vocales, solamente con sus primeras tres formantes y con la (reducida) base de datos PB.

Se desarrolla una transformación (lineal) dependiente del hablante a base de una red neuronal de segundo orden. Esta ANN se adapta a un nuevo hablante dejando fijo un clasificador posterior. La arquitectura utilizada tiene la particularidad de que, además de procesar las entradas normales de cualquier MLP (en este caso las formantes de las vocales), procesa todos los productos posibles entre las entradas mediante otro conjunto de pesos y las denominadas unidades de segundo orden. Otra característica que debe destacarse es que se trata de unidades lineales (aunque reciban productos de las entradas).

Las mejoras en la adaptación, en relación al sistema sin adaptar, son muy significativas (desde 78 a 95%) pero es necesario recordar lo limitado de su aplicación y generalización dada la base de datos utilizada. Entre los resultados se realiza una extensa comparación con otros métodos de normalización y adaptación al hablante.

Greenwood (1997) propone un algoritmo evolutivo para el entrenamiento de RNN parciales (PRNN). El sistema no está orientado al ASR en un sentido amplio sino que trata de la clasificación de visemas (grupos de fonemas que tienen similar movimiento de labios al pronunciarse). En principio este problema no se plantea como difícil debido a que los fonemas que se pronuncian con movimientos de labios similares son similares fonéticamente y sería más difícil distinguirlos entre ellos que distinguirlos en relación a los de otra clase. El sistema está orientado a que personas con problemas de audición que normalmente utilizan lectura labial como complemento, puedan utilizar el teléfono para comunicarse. Si el sistema funcionara bien, proveería a la persona esa ayuda de la lectura labial que no tiene al comunicarse telefónicamente.

Para cada visema se utiliza una PRNN separada que se entrena por medio de un algoritmo evolutivo. Luego se realiza la clasificación considerando como válida la red que mayor salida da. La

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

estructura de cada red consiste en 12 entradas, una salida de clasificación y cuatro salidas realimentadas como entradas. Para el entrenamiento se define una función de entropía relativa y se optimiza mediante un algoritmo evolutivo que tiene la particularidad de solamente utilizar mutaciones (dejando de lado un operador tan importante como la cruza). También es curioso que se utiliza una población muy pequeña (25 individuos). Los fonemas para el entrenamiento son extraídos de una región de la base de datos TIMIT. No se puede decir que los resultados sean buenos, aún considerando que el problema no es tan difícil. Cuando se compara con un entrenamiento basado en BP resulta ser que si bien el método evolutivo es más rápido[5], el de BP logra mejores resultados.

III. Nuevas arquitecturas neuronales para ASR

En esta sección se podrán encontrar trabajos mucho más originales e interesantes. En general se trata de nuevas arquitecturas que se diseñaron –casi exclusivamente– para el ASR.

III.1. ANN con comportamiento dinámico

Yamauchi y cols. (1995) proponen un modelo neuronal dinámico para la segmentación y clasificación de palabras independientemente de la velocidad con que se pronuncien. El sistema está compuesto de dos partes: la primera es un modelo de oído que dedica especial atención en la variaciones de entonación (en los lenguajes orientales, como es el caso, este parámetro es muy importante). La segunda parte es un modelo neuronal que tiene por objetivo realizar un reconocimiento independiente de las deformaciones introducidas por la velocidad de locución, pero sin dejar de reconocer las variaciones temporales que caracterizan a las palabras.

El modelo de oído se centra principalmente en la vía auditiva entre la cóclea y la corteza. Este modelo extrae parámetros relacionados con las variaciones de las frecuencias fundamentales del espectro de la voz. Se detectan regiones constantes y cambios ascendentes y descendentes en las componentes de frecuencia a lo largo del tiempo.

El reconocimiento lo realiza un sistema modular donde un bloque principal se encarga de la clasificación mientras otros dos bloques secundarios monitorean la velocidad de cambio en los patrones de entrada y controlan retardos temporales para acercar esta velocidad (vista desde el clasificador) a la velocidad que tenían los patrones de entrenamiento.

Los 3 módulos poseen la misma estructura: una línea de retardos temporales que alimentan a un MLP similar al neocognitrón. Lo que diferencia a cada estructura es la velocidad de propagación de las líneas de retardo. Uno de los bloques secundarios posee una línea de retardos más lenta que el bloque principal y el otro posee una línea de retardos más rápida. Todas las líneas de retardo responden a un controlador de velocidad que compara los promedios temporales de las salidas de los bloques secundarios y genera cambios para maximizar el promedio temporal de las salidas del bloque principal. Por ejemplo, si el promedio temporal de activación del bloque rápido es mayor que el del bloque lento, entonces el controlador aumenta las velocidades en las líneas de retardo del bloque principal. En lugar de tomar como salida final solamente la del bloque principal, las salidas de todos los bloques se suman con una ganancia individual que depende de sus respuestas.

Resumen para	(**Yamauchi** y cols., 1995)
Modelo	Estructura modular de tres MLP con diferentes líneas de retardos en el tiempo a la entrada. Un sistema controla la velocidad de estas líneas de retardo de acuerdo a las entradas y las salidas.
Entradas	Modelo de las vías auditivas propio, extracción de características de variación de las frecuencias principales
Salidas	Clasificación en tres palabras a lo largo del tiempo.
Entrenamiento	Se fijan los parámetros relacionados con la dinámica del sistema y se entrena por un algoritmo basado en “el-ganador-toma-todo” (no supervisado)
Experimentos	3 palabras y caracteres escritos
Base de datos	Propia, 2 hablantes en japonés
Reconocimiento	Palabras aisladas
Valoración de resultados	Regulares
Cobertura ASR	30%

Durante el entrenamiento todos los retardos permanecen fijos. En este sentido el enfoque es similar a muchos otros que no entrenan las propiedades dinámicas del sistema inteligente sino que las usan como un método de adaptación cuando los patrones de entrenamiento cambian en el tiempo. Se entrena un único neocognitrón por una regla no supervisada similar a la de “el-ganador-toma-todo”. Luego se copian todos los pesos a los otros bloques secundarios.

Para los experimentos se utilizaron tres palabras en japonés que se pronunciaron a velocidad constante para el entrenamiento y a velocidades variables para las pruebas. Entre cada palabra hay

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

un silencio importante. No se presentan resultados definitivos pero en lugar de eso hay una gráfica que muestra la evolución de las salidas en relación con las entradas. Dado el tipo de preprocesamiento y las palabras utilizadas parece ser que la entonación cumple un rol fundamental. A partir de los espectrogramas que se muestran en los resultados no parece difícil, incluso a simple vista, distinguir entre estas tres palabras.
También se presentan pruebas en reconocimiento de caracteres. Los autores destacan que este modelo es plausible biológicamente (en mamíferos superiores).
En (**Ceccarelli** y **Hounsou**, 1996), los autores utilizan redes con funciones de base radial (RBFN) modificadas para el reconocimiento de palabras aisladas. La modificación de las RBFN consiste en el agregado de unas conexiones con retardo y una capa de integración sobre el tiempo para capturar la naturaleza dinámica de la señal de voz. Los autores denominan el nuevo modelo como RBFN con retardos en el tiempo (TDRBFN). La red posee en tres capas, sin contar la de distribución. En la primera capa se encuentran las neuronas con activación RFB, en la segunda capa se encuentran las de activación sigmoidea y en la capa final se realiza una integración temporal de las salidas de la segunda capa. La particularidad incorporada es que entre la capa RBF y la siguiente, además de las conexiones normales, existen dos conjuntos de conexiones con retardos en el tiempo (de forma similar a una TDNN). La última capa, que también posee una función relacionada con el tiempo, simplemente suma las salidas de cada neurona de la capa anterior. Durante la clasificación, todos los vectores de características de una palabra son procesados por la red y la salida que mayor suma tiene al final es la que determina la palabra clasificada.
Los experimentos consisten en la comparación con diversas modalidades de entrenamiento y una TDNN. Para estas pruebas se utiliza una base de datos en la que se encuentran diez dígitos pronunciados, reiteradamente, por 60 hablantes italianos (28 para entrenamiento y 36 para las pruebas). Cada palabra es procesada para obtener 9 MFCC cada 20 ms. El vector de entrada para la red se forma con 10 vectores de características consecutivos.
En el primer experimento se calculan los centroides del núcleo radial mediante el entrenamiento no supervisado SOM y con una minimización del promedio de los errores cuadráticos (LMS) se ajustan los pesos de la capa sigmoidea. En el segundo experimento se cambia el SOM por el entrenamiento supervisado de cuantización vectorial con aprendizaje (LVQ). En el tercero se utiliza el algoritmo generalizado para RBFN (GRBF) con diferentes combinaciones de nodos en la capa RBF (la capa sigmoidea debe tener tantos nodos como palabras a clasificar). Finalmente, el último experimento consiste en la utilización de una TDNN. Los mejores resultados promedio para los experimentos son 94.8, 95.5, 95.1 y 89.8% respectivamente. No está claro si la diferencia entre el segundo y el tercero es significativa pero, en cuanto al costo computacional del entrenamiento, el método GRBF supera en un orden de magnitud al LVQ+LMS.

Resumen para	(**Ceccarelli** y **Hounsou**, 1996)
Modelo	TD-RBFN: una RBFN con retardos en el tiempo en sus conexiones.
Entradas	10 vectores consecutivos de 9 MFCC
Salidas	Se acumulan las activaciones de la capa sigmoidea para elegir la máxima y así clasificar.
Entrenamiento	SOM+LMS, LVQ+LMS y GRBF.
Experimentos	Distintos algoritmos de entrenamiento y comparación con TDNN
Base de datos	Propia, 64 hablantes italianos
Reconocimiento	Dígitos aislados
Valoración de resultados	Muy buenos
Cobertura ASR	40%

Nguyen y **Cottrell** (1997) proponen el modelo neuronal denominado *Tau Net*, especialmente ideado para modelar la variabilidad temporal en señales. En este modelo se usa una combinación de conexiones recurrentes, predictivas y con retardo en el tiempo. Para adaptarse a la variabilidad temporal de la señal, se adaptan unas constantes de tiempo de acuerdo al error de predicción.
La Tau Net utiliza un enfoque del tipo modelo neuronal predictivo (NPM) para el reconocimiento de secuencias. Según este enfoque, la red recibe un conjunto de vectores de características en el contexto del vector que tiene que dar como salida. En su estructura interna esta red posee una capa oculta completamente interconectada y con realimentaciones en todas sus neuronas. La capa de entrada (contexto) y la capa de salida (predicción) están completamente conectadas a la capa recurrente oculta. Tanto en la capa oculta como en la salida se utiliza la tangente hiperbólica como función de activación (la capa de entrada es simplemente de distribución). Cada nodo en la capa oculta posee una dinámica temporal que depende de una constante de tiempo (τ). En la fase de

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

entrenamiento la constante es fijada a un valor medio y los pesos son entrenados mediante gradiente descendiente sobre el error de predicción. Durante la fase de reconocimiento, los pesos ya quedan fijos y se adapta la constante de tiempo para reducir, nuevamente, el error de predicción (al aumentar τ la red responde más lentamente y a la inversa cuando se reduce τ).

Si bien es posible una constante τ por nodo en la capa oculta, finalmente los autores utilizan en las pruebas una sola constante para todos los nodos. Para los experimentos, la Tau Net es entrenada con señales de velocidad media y es probada con señales de todas las velocidades. La red posee la capacidad de extrapolar el reconocimiento a velocidades que no figuraban en el entrenamiento. Por otro lado, como en todo reconocimiento predictivo, se entrena una Tau Net para cada clase y luego se determina la clasificación eligiendo aquella Tau Net que posea menor error de predicción total para los vectores de entrada dados.

De la base de datos TIMIT se extraen 3 vocales (bien diferentes) y tres consonantes plosivas (no tan fáciles de clasificar). La clasificación que se realiza es independiente del hablante con una parametrización basada en 14 MSC.

Los resultados en reconocimiento de las tres vocales son muy buenos siendo notoria la diferencia en el desempeño con y sin adaptación de la constante de tiempo. Sin embargo, en el caso de las consonantes los errores son muy elevados y en varios casos el hecho de adaptar la constante de tiempo no redunda en ningún beneficio en relación al sistema sin adaptación.

Estos resultados probablemente se deban a que las vocales poseen una dinámica temporal más rica, generalmente caracterizada por una parte central estable que puede cambiar de duración con la velocidad del habla. Sin embargo las plosivas se presentan como un ligero transitorio y no poseen una parte estable donde pueda tener efecto el fenómeno de adaptación dinámica de la Tau Net. Este ligero transitorio de las plosivas no se ve sustancialmente modificado cuando se cambia la velocidad de elocución. Lo que naturalmente modificamos en estos casos es la zona estable de las vocales.

Finalmente hay que destacar que lo que se modela en esta red es más bien la velocidad con que cambian los patrones, y no necesariamente la forma con que éstos cambian. Es decir, no se prevé que al cambiar las velocidades también las características en sí mismas cambien y no sólo la velocidad con que se realizan los cambios. ¿Las características de un fonema en pronunciación rápida son las mismas que las del mismo fonema pronunciado lentamente? ¿Es equivalente a pasarlo en “cámara lenta” o a muestrearlo más rápido? La respuesta es no. Podría considerarse que estos cambios no modelados cabrían dentro de la capacidad de generalización del modelo neuronal pero entonces, también serían culpables de muchos de los errores de reconocimiento.

Resumen para	(**Nguyen** y **Cottrell**, 1997)
Modelo	Tau Net: ANN que funciona como NPM y ajusta su dinámica interna en función del error de predicción.
Entradas	Vectores de contexto compuestos por 14 MSC.
Salidas	Predicción del vector de características actual.
Entrenamiento	Gradiente descendiente sobre el error de predicción con la constante de tiempo fija.
Experimentos	3 vocales y 3 consonantes con y sin adaptación de la constante de tiempo
Base de datos	TIMIT (pre-segmentada)
Reconocimiento	Algunos fonemas
Valoración de resultados	Muy buenos para las vocales y muy malos para las plosivas
Cobertura ASR	20%

Djezzar y **Pican** (1997) utilizan un MLP con sensitividad contextual para el reconocimiento de 3 fonemas en francés. En el artículo se realiza un amplio estudio de las características fonológicas de los fonemas en cuestión y se obtiene un extractor de características.

El sistema de reconocimiento consiste en un MLP donde los pesos de las conexiones son adaptados dinámicamente en relación a las variaciones de los parámetros de contexto. Esta arquitectura se denomina: estimación ortogonal de variación de pesos (ODWE). Esta ODWE posee una estructura similar a un MLP y sus salidas se encargan de modificar los pesos de otro MLP (el clasificador propiamente dicho). En particular, dado que el trabajo está orientado al reconocimiento de tres consonantes, el sistema ODWE se encarga de modificar los pesos del MLP dependiendo la vocal que sigue a la consonante a clasificar. Para el entrenamiento se usa BP en una primera fase donde se entrena separadamente el MLP. En la segunda fase, también mediante BP, se entrena el ODWE dejando fijos los pesos del MLP.

En los experimentos se reconocen los fonemas /*p*/, /*t*/ y /*k*/, en francés, mediante la utilización del sistema con información contextual (ODWE-MLP) y mediante un MLP solo. También se realizan

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

algunos experimentos para encontrar el vector de características más apropiado. Los resultados con incorporación de la información contextual superan el 90%.

Resumen para	(**Djezzar** y **Pican**, 1997)
Modelo	ODWE-MLP: MLP en el que la información contextual modifica sus pesos.
Entradas	Procesamiento basado en características acústicas de los fonemas a reconocer.
Salidas	Clasificación de 3 fonemas.
Entrenamiento	BP en dos etapas.
Experimentos	Algunos experimentos para la parametrización, algunos sin contexto y otros con el sistema completo.
Base de datos	Propia, algunos fonemas.
Reconocimiento	Fonemas (/*p*/, /*t*/ y /*k*/) en contexto vocálico.
Valoración de resultados	Muy buenos.
Cobertura ASR	30%

III.2. ANN cercanas a un HMM

Un trabajo muy interesante es el de **Levin** (1993), donde se introduce la arquitectura de redes neuronales con control oculto (HCNN), especialmente diseñada para el modelado de series temporales. Esta muy particular ANN está inspirada en los HMM no sólo en su denominación sino a lo largo de todo su desarrollo. Se trata de una arquitectura de MLP que puede cambiar su funcionalidad mediante unas entradas adicionales de control. El entrenamiento está basado en BP y un mecanismo de segmentación para la estimación de las entradas de control oculto. Como resultado se obtiene un sistema de mapeo no-lineal y variante en el tiempo. Esta arquitectura neuronal es también descripta en términos probabilísticos como una generalización de un HMM.

Una HCNN es una ANN del tipo MLP que incorpora unas entradas adicionales que permiten que su mapeo se modifique sin necesidad de modificar los pesos. Si estas entradas adicionales para el control estuvieran fijas en el tiempo, entonces no habría distinción entre una HCNN y un MLP con más entradas. El caso es que la definición de las entradas de control es parte del algoritmo mismo de modelado y se encuentran “ocultas”, es decir, no son especificadas externamente sino que son generadas en el mismo sistema. Considerando que la entrada de control puede tomar sólo un conjunto finito de valores discretos, se puede analizar la HCNN como un autómata de estados finitos. Estados que quedan definidos en el tiempo mediante las entradas de control. Siguiendo la idea de los HMM en ASR, la autora restringe el análisis fundamentalmente a secuencias de estados de izquierda a derecha.

Luego el problema de la utilización de las HCNN se divide en tres puntos centrales: la segmentación, la evaluación y el entrenamiento. En cuanto a la segmentación se puede decir que se utiliza un método similar a la búsqueda de Viterbi p ara HMM: manteniendo los pesos fijos se busca cual es la secuencia de entradas de control que minimiza cierto criterio de error. Este criterio de error se define según la distancia euclídea entre la predicción de la HCNN y la salida correcta (nuevamente siguiendo a los HMM se valora la predicción de primer orden).

Una vez encontrada la secuencia óptima se puede pasar al problema de evaluación que consiste simplemente en conformar las entradas a la red como la concatenación del vector de entradas propiamente dicho y el vector de control (ya conocido). Se puede medir el error de predicción de la red nuevamente mediante la distancia euclídea acumulada entre la salida de la red y la salida deseada.

Para el problema de entrenamiento hay que considerar dos etapas iterativas de ajuste y segmentación. El ajuste se hace mediante el algoritmo de BP y la segmentación mediante una búsqueda similar a la de Viterbi.

Las pruebas se realizan en predicción de series temporales no-lineales y de tiempo variante (ecuación logística), reconocimiento de dígitos (hablados) y reconocimiento de caracteres escritos. En lo que interesa a este artículo hay que destacar que el reconocimiento es del tipo de palabras aisladas. Se realizó una segmentación previa en palabras con HMM y luego se aplicó las HCNN (que segmentaron los fonemas y reconocieron). La base de datos es propia, en línea telefónica y para hablantes masculinos. El preprocesamiento consiste en la obtención de 12 CC y sus 12 respectivos ΔC. Para cada palabra en el vocabulario (10 dígitos) se utilizó un modelo HCNN de derecha a izquierda con 8 estados. Los resultados indican un 99.1% de palabras correctamente reconocidas (no se obtienen valores de referencia para esta base de datos con HMM). Es de destacar que las segmentaciones comparadas entre un HMM y una HCNN son muy diferentes con lo que queda claro que el funcionamiento interno y las características capturadas por un modelo y el otro también lo son.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

Resumen para	(**Levin**, 1993)
Modelo	HCNN: arquitectura neuronal inspirada en un HMM.
Entradas	CC con sus ΔC
Salidas	Predicción de series temporales, método de clasificación por comparación entre diferentes modelos para cada palabra
Entrenamiento	Dos etapas: segmentación por un método basado en un algoritmo similar al de Viterbi y reestimación basada en BP
Experimentos	Mapeo de la ecuación logística, palabras aisladas y reconocimiento de caracteres escritos.
Base de datos	Propia, hablantes masculinos.
Reconocimiento	Palabras aisladas (10 dígitos)
Valoración de resultados	Muy buenos
Cobertura ASR	70%

Una extensión de estas ideas al ASR continuo de gran vocabulario es presentada en (**Petek** y cols., 1992). La estructura general del sistema está basada también en el concepto de NPM. Según este enfoque el modelo neuronal predice el próximo vector de características a partir de los vectores de contexto. Existen dos enfoques, uno desde las redes neuronales modulares (MNN) donde múltiples modelos compiten por el mejor modelado del próximo vector de características y el otro enfoque es el de las HCNN que mediante entradas de control emulan el comportamiento de diferentes MLP en el tiempo. En este último caso el costo computacional es mucho menor.

Los modelos de palabras se forman mediante la concatenación de varias HCNN de fonemas (de forma similar a como se hace con los HMM). Esto es beneficioso para tratar el problema de reconocimiento en gran vocabulario. Sin embargo, también se realizaron pruebas modelando algunas palabras clave (“a”, “the” y “you”) mediante una única HCNN. También se consideraron otras mejoras como el modelado de fonemas en contexto, poniendo así al sistema basado en HCNN a la altura de los mejores sistemas HMM de la época (lamentablemente no se proveen resultados comparativos con un sistema HMM para exactamente la misma tarea de reconocimiento). Las pruebas se realizaron para un solo hablante lo cuál quizás constituya uno de lo pocos puntos débiles del trabajo.

Resumen para	(**Petek** y cols., 1992)
Modelo	HCNN
Entradas	MFCC con contexto.
Salidas	Predicción del próximo vector de voz, modelos de fonemas conectados
Entrenamiento	Extensión de HCNN para modelos concatenados.
Experimentos	Reconocimiento de voz continua con una MNN y dos variantes de HCNN.
Base de datos	Un único hablante, 204 frases de CMU's Conference Registration Database.
Reconocimiento	Habla continua pero un solo hablante
Valoración de resultados	Buenos
Cobertura ASR	80%

Otra combinación entre programación dinámica (DP) y MLP es la propuesta por **Martens** y **Depuydt** (1991). En este caso se utiliza el híbrido para la clasificación de fonemas según sus características articulatorias y números conectados en holandés. En cuanto a la clasificación de fonemas según sus características fonéticas, la novedad es que a un MLP se lo alimenta con diferentes combinaciones del vector actual de características y otros de contexto. La clasificación fue hecha sobre una base de datos propia, segmentada manualmente. El MLP se entrenó mediante BP para diferentes números de neuronas en la capa oculta. Si bien los resultados son buenos (85.3% de error con 30 nodos en la capa oculta) hay que recordar que no se trata de una clasificación fonética definitiva sino solamente según sus tipos de articulación.

El algoritmo de segmentación se basa en las salidas del MLP para, mediante agrupación y eliminación de espurios con DP, separar la palabra en fonemas. Los autores evalúan el desempeño de su algoritmo para la clasificación de sólo 3 tipos articulatorios. De esta forma pueden comparar su algoritmo con otros previamente publicados. Los resultados son algo inferiores pero la segmentación puede superar a la realizada por modelos HMM (no se especifica de que tipo pero la referencia citada es de 1989).

III.3. Hacia una unificación de los paradigmas

En el trabajo presentado por **Bridle** (1990) se encuentra un modelo conexionista que, tanto en estructura formal como en método de entrenamiento, puede ser interpretado como un HMM. El desarrollo del artículo está restringido a un enfoque de reconocimiento de palabras aisladas (donde el inicio y el fin de palabra ya han sido identificados) aunque es extensible a otras modalidades. Se

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

plantea la obtención de los coeficientes α (típicos en el cálculo de las probabilidades de un HMM) mediante una red recurrente (Alpha-net). Esta semejanza se desarrolla en una arquitectura algo particular que quizás no llega ser neuronal. A continuación se puede expandir esta red recurrente como un MLP con tantas capas como vectores de voz contenga la palabra a reconocer. Finalmente se logra establecer un paralelo entre la retropropagación de un MLP y la pasada hacia atrás del método de Baum-Welch para el entrenamiento de los HMM.

Un detalle en este trabajo es que se une a aquellos que denominan "red neuronal" a una arquitectura que difícilmente posea todas las características con las que se define comúnmente a una red neuronal. De hecho, en este caso se trata de una red o modelo conexionista y el autor mismo reconoce que no es muy neuronal cerrando la palabra 'neuronal' en comillas simples.

Como posteriormente destacarían otros autores, no queda claro qué ventajas prácticas tiene esta interpretación en relación a la clásica de los HMM (porque, a fin de cuentas, son lo mismo). Más aún, este modelo poseería las mismas desventajas o falencias de los HMM. Eso sí, más allá de lo interesante del trazo formal del paralelismo paradigmático, queda por esperar que esta nueva interpretación abra una vía conceptual diferente para la obtención de nuevos sistemas de ASR o un mejoramiento sustancial de los ya existentes.

Resumen para	(**Bridle**, 1990)
Modelo	Red recurrente expandida en MLP.
Entradas	Cualquiera de las utilizadas normalmente para ASR.
Salidas	Probabilidad acumulada equivalente a la de un HMM.
Entrenamiento	Gradiente descendiente
Experimentos	No se realizan.
Base de datos	No se usa.
Reconocimiento	Palabras aisladas (extensible).
Valoración de resultados	Muy buenos
Cobertura ASR	50%

A pesar de no realizarse experimentos específicamente de ASR, se ha decidido incluir un breve comentario acerca del trabajo de **Bengio** y **Frasconi** (1996), debido a la cercanía de los problemas que trata con el ASR y la gran cantidad de citas desde otros artículos. En este trabajo se propone un modelo neuronal híbrido, con interpretación estadística desde los HMM, denominado "modelo ocultos de Markov de entrada-salida" (IOHMM). En este trabajo se considera el modelado de secuencias temporales mediante una arquitectura neuronal recurrente modular que contempla una ANN para cada estado de un HMM en el tiempo. Se propone un método para el entrenamiento que es valorado como poseedor de un poder discriminativo superior a cualquier algoritmo de entrenamiento de HMM.

El método consiste en dos grandes etapas. En la primera se aproxima la predicción de la secuencia interna de estados mediante pseudoobjetivos y en la segunda se ajustan los parámetros para acercar los pseudoobjetivos de trayectorias de estados a los objetivos de salida requeridos. De no existir conocimiento a priori, se inicializan las trayectorias de pseudoobjetivos al azar y el método de optimización asegura que el sistema llegue a los verdaderos objetivos.

En particular hay que notar que la arquitectura tiene más particularidades de un HMM que de un modelo conexionista o MNN. Por ejemplo, las variables de estado son discretas y los elementos de procesamiento son lineales. El método de entrenamiento está inspirado en la maximización de la esperanza (común a los HMM). La principal diferencia entre los clásicos HMM y los propuestos IOHMM es que, mientras los HMM representan la distribución de probabilidades para determinadas secuencias de salidas, los IOHMM representan estas probabilidades pero condicionadas a determinadas secuencias de entradas.

Los resultados están orientados a la inferencia de gramáticas y, en particular, los experimentos tratan con las 7 gramáticas de Tomita.

III.4. Sistemas modulares

Otro trabajo en el que se aplican las HCNN de Levin es el presentado por **Kim** y **Lee** en 1994. En un primer nivel se encuentran un MLP y una HCNN para realizar una preclasificación que es "juzgada" en un segundo nivel por un sistema inteligente. De aquí se deriva el nombre que se le dio al modelo: red neuronal de juicio inteligente (IJNN). Para realizar este juicio inteligente se prueban dos alternativas: un sistema neuronal (*neural-judge*) y un sistema difuso (*fuzzy-judge*).

El MLP clasifica directamente en palabras. Existe una previa normalización en el tiempo que elimina los vectores de características similares hasta lograr un determinado número (fijo) de vectores que

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

analiza el MLP. La HCNN trabaja como un NPM y existen tantas HCNN como palabras a modelar. Debido a que es poco probable que los dos clasificadores se confundan al mismo tiempo (en opinión de los autores) el tercer módulo puede ser aplicado para cuando exista una disputa entre ambos.
Los módulos del nivel inferior son entrenados independientemente (por BP y error de predicción en las HCNN). En el caso del nivel superior con MLP, éste es entrenado por BP con los mismos datos que los clasificadores del nivel inferior[6].
Los experimentos se realizan en reconocimiento de dígitos aislados en coreano. La base de datos se obtiene a partir de 6 hablantes para entrenamiento y 6 para prueba. El preprocesamiento consiste en la obtención de 14 CC y ΔC de energía. Los resultados muestran una mejora significativa en relación a la HCNN sola pero no está claro cuan significativas son las diferencias entre los otros métodos (MLP solo, IJNN neuronal e IJNN difusa).
Cosi y cols. (1996) proponen un sistema neuronal de integración de información auditiva y visual para el reconocimiento de consonantes plosivas del italiano. Basándose en la ayuda que puede otorgar la lectura labial, se integra mediante dos RNN el conocimiento que proviene de las características de la voz y de un sistema especializado para la detección de los movimientos labiales.
Las RNN poseen nodos recurrentes sólo en la capa oculta. Estos nodos recurrentes tienen una realimentación de sus salidas a si mismos y cada entrada al nodo (incluyendo la realimentación) tiene 3 retardos temporales. No se brindan detalles del algoritmo de entrenamiento pero se trata de una extensión de BP para redes recurrentes (retropropagación para secuencias). Se realizan experimentos tanto para un sistema dependiente del hablante como para otro en el que se entrena con 9 hablantes y se reconoce con uno diferente. Para el caso dependiente del hablante se logran tasas de reconocimientos siempre mayores al 90% (consonantes plosivas italianas). En algunos de los casos independientes del hablante se supera el 90% pero en otros no se llega al 50%.
Otra estructura modular se basada en TDNN se propone para el reconocimiento de la parte final de todas las sílabas del mandarín (**Poo**, 1997). La arquitectura se denomina TDNN de dos niveles (TLTDNN). En el primer nivel se discrimina entre el grupo de las vocales y el grupo de las consonantes nasales. En el segundo nivel se realiza la clasificación en los 35 posibles fonemas de la parte final de las sílabas del mandarín. El primer nivel está compuesto por 2 subredes, una encargada de los subgrupos de vocales y otra encargada del grupo de la vocal /*a*/, que tiene diferentes terminaciones nasales. El segundo nivel se compone por 8 redes pequeñas que se encargan de discriminar a partir de los subgrupos clasificados en el nivel 1. Todas las subredes se entrenan independientemente por BP a través del tiempo (BPTT).
La base de datos consiste en 8 conjuntos de 1265 monosílabos hablados por una hablante femenina. Todas estas elocuciones fueron segmentadas manualmente en fonemas para el entrenamiento y las pruebas.
Los resultados incluyen los desempeños para cada nivel independientemente y para el sistema completo. Los resultados finales alcanzan el 95.6%.

Resumen para	(**Poo**, 1997)
Modelo	Modular en dos niveles (TLTDNN)
Entradas	16 MFCC con 2 vectores de contexto
Salidas	Fonemas de la parte final de las sílabas del mandarín.
Entrenamiento	BPTT
Experimentos	Diferentes módulos separadamente y sistema completo
Base de datos	Propia, un hablante femenino, segmentada manualmente.
Reconocimiento	Algunos fonemas
Valoración de resultados	Muy buenos
Cobertura ASR	20%

En (**Chudý** y cols., 1991) se presenta un sistema totalmente neuronal, inspirado en el sistema de percepción de los loros, para el reconocimiento de palabras aisladas (en eslovaco). El sistema puede ser dividido en dos partes: un preprocesador de percepción neuronal y una red neuronal probabilística (PNN). El preprocesador consiste en un modelo de oído donde se pueden diferenciar: oído medio, membrana basilar, células ciliadas, nervio auditivo y neuronas de la vía auditiva. Se describe cada una de las etapas y se analizan en términos de compresión de la información. En cuanto a la PNN los autores obtienen una función discriminante basada en un análisis probabilístico y proponen un MLP para la aproximación. No se trata de un MLP en el sentido más clásico, éste posee una capa de entrada (de distribución), una capa intermedia con una no-linealidad de tipo

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

gaussiana, una capa de sumación (nodos lineales) y al final una MAXNET encargada de devolver el índice del nodo cuya suma sea mayor. Esta red es entrenada por BP.
Para los experimentos se utilizan solamente 5 palabras que difieren en la vocal final. La base de datos se registró a partir de 5 hablantes masculinos y 3 femeninos. En el preprocesamiento se obtiene un vector de características de 32 elementos. Los resultados van desde el 58 al 98% con un promedio de 89%.
En el trabajo de **Chen** y **Liao** (1998) se utiliza una arquitectura modular de redes neuronales recurrentes para el reconocimiento de 1280 sílabas del mandarín. La modularidad del sistema está definida por un conocimiento a priori de las características fonéticas y lingüísticas de este lenguaje. Si bien se considera la dinámica temporal en el modelado, el reconocimiento no llega a ser continuo. Se trata del modelado de una dinámica a corto tiempo que sirve para un reconocimiento de palabras aisladas (sílabas en este caso) y no puede ser extendido directamente a un sistema de reconocimiento de habla continua.
Hay que destacar que los lenguajes orientales tienen importantes particularidades que hacen que las técnicas utilizadas para el ASR sean poco extrapolables a los lenguajes occidentales (y viceversa). Por lo tanto en este artículo, donde diferentes RNN se encargan de atacar las distintas características de las sílabas del mandarín, quizás sean pocas las ideas que se puedan extraer para un sistema de ASR de lenguajes occidentales.
La primera característica que es atacada por una RNN separada es la entonación. En los lenguajes orientales la entonación puede cambiar totalmente la semántica de una palabra y por lo tanto se justifica la existencia de una RNN que procesa la curva de entonación y clasifica en 5 posibles movimientos de este parámetro. Otras dos RNN se encargan de procesar las partes inicial y final de las sílabas que, en general, se trata de consonantes y vocal-consonantes respectivamente. Estas redes trabajan con salidas que indican parámetros fonéticos como el lugar y la forma de la articulación. Finalmente existen dos RNN más que se encargan de “pesar” las salidas de las otras tres RNN para que un módulo posterior realice un proceso de acumulación de funciones discriminantes para tomar la decisión.
Las 5 RNN poseen una estructura de 3 capas y todas las salidas de las capas ocultas son realimentadas como entradas a sí mismas. Las salidas de las RNN son lineales y las tres primeras se entrenan mediante el algoritmo de minimización del error de clasificación denominado: probabilidad descendiente generalizada. Las dos RNN de pesado se entrenan mediante BP. Sin embargo hay que destacar que para el entrenamiento se requiere una previa segmentación en sílabas y entre la parte inicial y final de las sílabas. Los autores la realizan con un método basado en alineación de HMM. No se requiere la segmentación entre la parte inicial y final de las sílabas cuando el sistema de reconocimiento ya está entrenado pero sigue siendo necesaria la segmentación externa en palabras.
Las pruebas se realizaron con una base de datos propia con 8 hablantes masculinos y 2 femeninos.
Las características utilizadas para las redes de reconocimiento de sílabas fueron 14 CC con sus ΔC, la primera y segunda derivada de la energía en el tiempo y la frecuencia de cruces por cero. Para la RNN de entonación se utilizó la energía, el ΔC de energía, el pico de la función de autocorrelación normalizada, la entonación y el ΔC de entonación.
Los resultados finales se indican para diferentes cantidades de neuronas en las capas ocultas y diferentes variantes y combinaciones del algoritmo de entrenamiento. También se muestran resultados con sistemas basados en HMM para el contraste. Para el mismo número de parámetros, los resultados del método propuesto son muy superiores a los obtenidos con HMM. Queda pendiente la extensión de la técnica para reconocimiento de habla continua.

Resumen para	(**Chen** y **Liao**, 1998)
Modelo	5 RNN que modelan diferentes aspectos fonéticos de las sílabas.
Entradas	Para 4 RNN se usan CC, con sus ΔC y energía. Para la quinta RNN se utilizan características relacionadas con la entonación.
Salidas	2 redes para el lugar y manera de articulación, otra para la modalidad de entonación y otras 2 para los pesos para ponderar las anteriores. Un sistema no neuronal las combina al final para tomar la decisión acerca de qué sílaba se trata.
Entrenamiento	Métodos combinados. Requiere segmentación por HMM.
Experimentos	Comparación con métodos basados en HMM y diferentes algoritmos de entrenamiento.
Base de datos	Propia, 8 hablantes masculinos y 2 femeninos.
Reconocimiento	Palabras aisladas: 1280 sílabas del mandarín.

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

Valoración de resultados	Muy buenos para igual número de parámetros en relación a otros paradigmas.
Cobertura ASR	50%

Lee y **Ching** (1999) presentan un sistema de reconocimiento de sílabas aisladas del cantonés, basado en un MLP y varias RNN que convergen en un sistema de integración final. El MLP se encarga del reconocimiento de las cadencias tonales. Las RNN se encargan del reconocimiento de las características fonéticas de las sílabas modelando a cada una separadamente (una RNN para cada sílaba). Como es común en los enfoques modulares, se destaca que la reconfiguración para incorporar nuevas sílabas al sistema es mínima, pero, en el mismo sentido, no se considera la reducción del poder discriminativo que esto genera. El algoritmo de integración final selecciona y elimina los modelos de sílabas menos probables a través de múltiples pasadas.

Hay que destacar que, nuevamente, al tratarse de un lenguaje oriental, es difícil hacer extrapolaciones de los enfoques y resultados hacia los lenguajes alfabéticos occidentales. El cantonés, de forma similar que el mandarín, es un lenguaje básicamente monosilábico y tonal.

El MLP encargado del reconocimiento tonal recibe 16 características de la curva de entonación de una sílaba, entre ellas: entonación inicial y final, cadencias, duración y energía relativa a la duración.

El entrenamiento está basado en BP y se realizaron pruebas en las que había desde 25 a 35 neuronas en la capa intermedia. Hay que destacar que los autores utilizan una normalización con respecto al hablante y esto hace que, al menos en un punto, el sistema final quede dependiente del hablante.

Los modelos para cada sílaba se logran con una gran cantidad de pequeñas RNN (12 neuronas) completamente conectadas. De las 12 neuronas 5 son consideradas como salidas, representando 5 partes de la sílaba. Partiendo de las características espectrales de cada sílaba el entrenamiento consta de dos etapas. En la primera se entrena cada RNN independientemente para generar la secuencia de estados apropiada a partir de los vectores de características de la sílaba que esta red modela. La segunda etapa consiste en un refinamiento global para mejorar la capacidad discriminativa.

Resumen para	(**Lee** y **Ching**, 1999)
Modelo	MLP-RNN.
Entradas	SC y diversas características de la entonación.
Salidas	RNN independientes para cada sílaba y MLP para clasificación de cadencias tonales.
Entrenamiento	MLP por BP y RNN en dos etapas, una independiente y la segunda de optimización global para incrementar la capacidad discriminativa.
Experimentos	Desempeño de las ANN por separado y en el sistema integrado para diferentes cantidades de sílabas.
Base de datos	Propia, 200 sílabas aisladas.
Reconocimiento	Sílabas aisladas del cantonés.
Valoración de resultados	Buenos
Cobertura ASR	60%

Se describen varias pruebas en el artículo. Primeramente se evalúa el clasificador de cadencias tonales, que alcanza un desempeño equivalente al alcanzado por oyentes humanos en condiciones similares. Luego se evalúa el sistema con 40 tipos de sílabas (independientes de la cadencia tonal) y con una expansión a 80 tipos. Los resultados superan el 90% sobre una base de datos propia que representa aproximadamente un 39% del contenido de un periódico local. Todos los resultados se obtienen para sílabas aisladas por lo que el método no realiza el proceso de segmentación en sílabas y no es aplicable al habla continua. Los resultados para 120 sílabas caen al 88% y en la expansión a 200 sílabas caen hasta el 81%.

IV. Combinaciones de ANN y HMM

Entre los variados enfoques para la combinación de ANN y HMM se encuentran muchos trabajos en dos modalidades básicas. En un primer grupo se tiene a las ANN para calcular las probabilidades de los HMM y en un segundo grupo están los trabajos en los que se utiliza una ANN para obtener las que serán entradas de un HMM. Son estos los dos primeros criterios que se han seguido para separar los trabajos de esta sección. Dentro de cada subsección los artículos se describen por orden cronológico.

IV.1. ANNs para la estimación de probabilidades de HMM

Resulta apropiado comenzar este apartado con un breve resumen de los trabajos precursores de Bourlard y cols. Difícilmente se puedan abarcar todos sus trabajos en el área pero los consideraremos una selección agrupada como buena introducción a la utilización de ANN para la estimación de probabilidades de HMM.

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

En uno de los primeros artículos en el tema, (**Bourlard** y **Wellekens**, 1990), los autores proponen la utilización de un tipo particular de MLP para la estimación de las probabilidades de observación de un HMM. Una de las principales ventajas de estos enfoques es que la información contextual se puede incorporar fácilmente y sin incrementar significativamente la complejidad del sistema. Esto no ocurre en un sistema HMM estándar ya que se obtiene un aumento de los parámetros a entrenar que es proporcional al cuadrado de los elementos que se agregan y así también aumenta la cantidad de datos necesarios para el entrenamiento. La segunda ventaja es un aumento de la capacidad discriminativa del sistema, en relación a un HMM estándar. La tercera ventaja es la capacidad de interpolación del MLP cuando existen pocos datos para el entrenamiento.

Se considera un MLP que recibe como entrada el índice (en codificación binaria) del prototipo de cuantización vectorial que normalmente alimenta a un HMM discreto (DHMM) o bien, directamente, el vector de características. Como salida el MLP debe dar la clase (palabra, fonema o estado) a la que pertenece dicha entrada. El MLP deberá realizar este mapeo y puede ser entrenado por gradiente descendiente con una base de datos previamente segmentada o bien conjuntamente con el algoritmo de segmentación de un HMM. Para este MLP se han probado varias configuraciones: una con vectores de contexto en la entrada, otra con realimentación de las salidas hacia la entrada (como una red de Jordan) y una tercera con realimentación y contexto. Para solucionar el problema de que las salidas no pueden ser interpretadas como probabilidades (locales) porque no suman 1, se utiliza la función exponencial normalizada ("softmax") en lugar de la sigmoidea. Así, para cada vector de entrada podemos obtener mediante el MLP la probabilidad de pertenencia a cada una de las clases que interese modelar en el HMM.

En los experimentos se prueba el MLP para clasificación fonética a nivel de vectores de entrada y para estimar las probabilidades de emisión de un modelo DHMM. En estos experimentos se utiliza la base de datos SPICOS, en alemán con un solo hablante y unas 200 frases. Se utilizaron para la entrada 30 MSC obtenidos cada 10 ms. Estos vectores son cuantizados según 132 centroides y la entrada del MLP se completa con 8 vectores de contexto. Para la capa intermedia se probaron entre 20 y 200 neuronas y en la capa de salida había 50 neuronas correspondientes a los 50 fonemas a clasificar. Para la clasificación fonética a nivel de vectores de características se logró un 62.7% en el mejor de los casos y se comparó con otros métodos estadísticos de clasificación. También se muestran comparativas para un MLP sin entradas de contexto. Para la base de datos TIMIT se logró solamente 40.9% (aunque superior a otros métodos de reconocimiento de patrones).

En la segunda serie de experimentos se utiliza un MLP como estimador de las probabilidades de un HMM. Este método permite que el sistema reconozca voz continua. En este caso el MLP es también entrenado para la clasificación en fonemas y sus salidas son utilizadas en el algoritmo de búsqueda por DP. En un primer experimento las salidas del MLP se utilizan como probabilidades de transición. En el segundo experimento son utilizadas como probabilidades de emisión. No se muestran en este trabajo resultados definitivos para estas estructuras. En 1992 se publica una generalización de estas ideas para incorporar modelos de trifonos (**Bourlard** y cols., 1992) y otros experimentos como, por ejemplo, los relativos a la utilización de RBFN en el modelo ANN-HMM. En resultados posteriormente publicados los sistemas híbridos MLP-HMM superan a los sistemas HMM clásicos. Por ejemplo, en el artículo (**Bourlard** y **Morgan**, 1993) se presentan errores de reconocimiento en palabras, donde se pueden observar importantes mejoras por parte de los sistemas híbridos (con MLP independientes del contexto). También en este artículo se puede encontrar una interesante revisión de ANN aplicadas a ASR (principalmente en combinación con HMM). Se destaca sobre el final un análisis comparativo de los costos computacionales, requerimientos de memoria y accesos a memoria para sistemas basados en HMM y MLP-HMM.

En el amplio trabajo de revisión y discusión (**Bourlard** y cols., 1996), se insta a los lectores a encontrar nuevas soluciones al problema de ASR aunque esto implique, inicialmente, incrementar los errores en relación a los sistemas actuales. En este trabajo se sostiene que por diversas razones (principalmente restricciones impuestas por las entidades que financian los trabajos) las investigaciones actuales relacionadas con el ASR están estancadas en un mínimo local y se necesitan nuevas ideas para escapar de allí. Sin embargo, estas nuevas ideas, quizás radicalmente diferentes, no tienen porque ignorar todo el conocimiento actual sobre el tema y difícilmente surjan de la nada.

Los trabajos de Bourlard y sus colaboradores han sido precursores de casi todos los trabajos que se incluyen en las siguientes secciones de este artículo. Aún más, también fuera del área de ASR se

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

pueden encontrar trabajos que son continuación de estas ideas (por ejemplo el **Baldi** y **Chauvin** (1996) para el modelado de la familia de las inmunoglobulinas).

Robinson (1994) presenta un muy interesante trabajo donde se utilizan RNN para la estimación de probabilidades de fonemas en sistemas de reconocimiento de gran vocabulario. Se puede considerar que este trabajo sigue la línea de Bourlard y cols., sólo que en este caso el autor ataca el problema de modelización del contexto a corto tiempo y los efectos de la coarticulación a través de una RNN. El problema en el caso de los HMM estándar es que esta extensión genera un incremento exponencial del número de modelos que suele ser prohibitivo en muchas aplicaciones reales. Aún más, el aumento de los parámetros en la matriz de covarianza de cada clase en un HMM es proporcional al cuadrado del aumento en la dimensionalidad de los vectores de entradas, lo cual, nuevamente limita la capacidad de expresión del modelado.

Se pueden observar dos extremos en el enfoque de híbridos ANN-HMM: un modelo conexionista donde se tienen tantas capas como segmentos de adquisición existan en la señal de voz a analizar (Alpha-nets) y, por otro lado, estimadores de probabilidades de fonemas que se entrenan independientemente de la estimación de las probabilidades de transición en los HMM. En el trabajo de Robinson se utiliza una RNN para la estimación de las probabilidades de fonemas que luego se incorporan en un HMM para el modelado de palabras. Se trata de un MLP con realimentaciones que permiten considerar el contexto a largo plazo. Las entradas se componen por la concatenación del vector de características de voz y un vector con un subgrupo de las salidas en el instante anterior. Sin embargo hay que destacar que las salidas realimentadas no forman parte del conjunto de salidas que se utilizan para la estimación de probabilidades. Aún más, el autor separa los pesos de la RNN en dos matrices, una para las salidas que se utilizarán para la estimación de probabilidades y otra para las salidas que serán realimentadas. Para el entrenamiento se realiza una expansión de la RNN en un número finito de instantes de tiempo. Si este intervalo de tiempo es suficientemente largo en relación a la duración de los transitorios que quieren capturarse, la red podrá modelar las características dinámicas que se necesitan. En particular, en el caso de la voz, se puede realizar una estimación de este intervalo de interés basándose en conocimientos fonológicos. Luego, se aplica una extensión del algoritmo de BP (en particular BPTT). Otra particularidad de la arquitectura neuronal es que las neuronas de salida no utilizan la clásica función sigmoidea sino que ésta se reemplaza por la “softmax”.

Se ha utilizado la base de datos TIMIT para el reconocimiento de fonemas y, para el reconocimiento de palabras aisladas, se utilizó la base de datos DARPA-1000 (si bien con una baja perplejidad, es meritorio para la época). Las características de la voz utilizadas fueron: la energía, la frecuencia fundamental, el grado de sonoridad y 20 MSC normalizados.

La salida de la RNN es considerada como probabilidad *a posteriori* de la clase de fonema, dada la evidencia acústica. Mediante la aplicación de la regla de Bayes, es posible convertir esta probabilidad en la probabilidad escalada de la evidencia acústica dada la clase de fonema (como requieren los HMM). Esta probabilidad se utiliza en los HMM en lugar de la estimación calculada a partir de las mezclas de gaussianas. Se utiliza un HMM muy simple (con todas las transiciones hacia adelante permitidas) y el algoritmo estándar de Viterbi para encontrar la secuencia de estados más probable.

Los resultados para el reconocimiento de fonemas se presentan ampliamente contrastados con otros reconocedores basados en diferentes paradigmas. Hay que destacar que el reconocedor de fonemas con RNN-HMM no supera el reconocimiento de un sistema basado solamente en HMM con el conjunto de herramientas *HTK*[7]. Sin embargo las diferencias seguramente no sean estadísticamente significativas (cosa que el autor no analiza) ya que se trata de un punto en porcentajes de reconocimiento que oscilan en torno al 75%.

En el caso del reconocimiento de palabras los resultados parecen mejores (95%) pero no hay comparativas para tomar como referencia (el autor dice que son 10% mejores a otros publicados anteriormente).

Por último cabe destacar que en este tipo de trabajos, si bien se abarca el problema del ASR casi en su totalidad, sigue siendo el paradigma del los HMM el que predomina y las ANN sólo mejoran la estimación de algún grupo de parámetros.

Resumen para	(**Robinson**, 1994)
Modelo	RNN: MLP con realimentación parcial estima las probabilidades de un sistema HMM.
Entradas	Energía, entonación, sonoridad y 20 MSC
Salidas	Neuronas “softmax” cuyas salidas son utilizadas como probabilidades

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

Entrenamiento	BPTT con algunas optimizaciones, término de momento y programación para 64 procesadores paralelos.
Experimentos	Reconocimiento de fonemas (bien contrastado con otros paradigmas) y de palabras (falta contraste)
Base de datos	TIMIT y DARPA-1000
Reconocimiento	En habla continua, vocabulario de tamaño medio (grande para la época).
Valoración de resultados	Buenos
Cobertura ASR	85%

En el trabajo de **Yan** (1999) se presenta el método de entrenamiento de un MLP utilizado para la estimación de probabilidades de observación de fonemas de un sistema basado en HMM. Este método reemplazaría la clásica codificación binaria de la salida deseada, donde para cada patrón de entrada se exige que sólo una salida (correspondiente a la clase etiquetada) sea uno. La correlación entre las salidas de una ANN puede ser vista como una medida de su similitud acústica.

Inicialmente el MLP es entrenado con salidas binarias y luego se utilizan estas salidas sobre toda la base de datos para obtener los coeficientes de correlación. El entrenamiento continúa de forma que para cada patrón de entrenamiento la salida deseada de la neurona correcta es 1.0 y las demás se igualan a su coeficiente de correlación con la salida correcta (similitud con la salida correcta), con un factor de escala determinado empíricamente.

Para el modelado de los fonemas se utilizó un sistema HMM con fonemas de contexto (trifonos). Cada estado del modelo de un fonema se corresponde con una salida del MLP. La señal de voz se procesa de forma de obtener 12 MFCC, la energía normalizada y sus ΔC. Como patrones de entrenamiento se concatenan 5 vectores de características (es decir 130 nodos de entrada para el MLP). La capa oculta posee 200 neuronas y en la salida hay 534.

Para el caso de palabras aisladas se realizaron tres experimentos con la base de datos PhoneBook. El error de referencia con el sistema ANN-HMM con el MLP entrenado con salidas binarias fue de 16%. Cuando se utilizó el MLP con el nuevo método de entrenamiento se logró un error de 12.7%. Finalmente se utilizó un sistema de referencia basado en HMM, implementado con el conjunto de herramientas *HTK* (no se usaban fonemas de contexto en este sistema de referencia). En este caso se llegó a un error del 12% pero el sistema ANN-HMM poseía sólo un 13% de los parámetros que poseía el de HMM.

Las pruebas de dígitos conectados se realizaron con la base de datos OGI_30K. En este caso sólo se comparan resultados entre los dos métodos de entrenamiento del MLP, utilizando la ANN entrenada con salidas binarias como inicialización para entrenar la ANN con el método propuesto (cosa que pone en clara desventaja a la primera). Como resultado se obtiene una reducción porcentual del error del 13% (desde 16% a 13.6%).

Resumen para	(**Yan**, 1999)
Modelo	MLP para la estimación de probabilidades de un sistema HMM.
Entradas	12 MFCC, energía, sus ΔC y 4 vectores de contexto.
Salidas	Estimación de las probabilidades de observación de un modelo HMM basado en trifonos
Entrenamiento	Dos etapas: entrenamiento con salidas deseadas binarias y luego con las basadas en la correlación de las salidas binarias
Experimentos	Palabras aisladas con MLPbinario-HMM, MLPcorrel-HMM y HMM puro. Dígitos conectados con MLPbinario-HMM y MLPcorrel-HMM.
Base de datos	PhoneBook y OGI_30K.
Reconocimiento	Palabras aisladas y dígitos conectados
Valoración de resultados	Buenos
Cobertura ASR	60%

Un trabajo muy interesante y más reciente es el de **Krogh** y **Riis** (1999), donde se describe un modelo denominado "redes neuronales ocultas" (HNN). El punto de partida de este trabajo es un modelo denominado class -HMM[8] (de los mismos autores). Estos modelos consisten en un HMM con una distribución de probabilidades sobre las clases asignadas a cada estado. Para mejorar la capacidad discriminativa del modelo, durante el entrenamiento se supone que cada estado describe una determinada clase. Entonces cada estado puede poseer una distribución de probabilidad para cada clase o etiqueta (de forma similar a los pseudoobjetivos de los IOHMM). El entrenamiento se realiza en dos fases: en la primera se entrena con cada estado fijo a una clase y se tienen en cuenta sólo los caminos cuya secuencia sea igual a la secuencia de entrenamiento; en la segunda fase esto se relaja de forma que las etiquetas de clase en los estados no se tienen en cuenta. En esta segunda etapa, se trabaja de forma similar a la decodificación estándar de un HMM y sirve también

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

para el reconocimiento mediante los class-HMM. Sin embargo, las restricciones impuestas en la primera etapa hacen que el algoritmo clásico de Baum-Welch no sea aplicable y los autores describen otro algoritmo basado en el descenso por el gradiente de una función de probabilidad compuesta por las probabilidades logarítmicas del modelo completo en las dos etapas. En el artículo también se provee una interpretación del modelo class-HMM según las redes de independencia probabilística (PIN). Este enfoque provee un método gráfico para comparar el funcionamiento de este modelo con otros (en términos de dependencias probabilísticas). En (**Smyth** y cols., 1997) se puede encontrar una muy completa revisión de PIN para HMM.

Para el híbrido entre class-HMM y ANN, denominado HNN, las probabilidades de emisión, de transición y de clase son reemplazadas por las salidas de tres MLP por cada estado, que toman los vectores de características con contexto como entrada. Todas las redes utilizan función sigmoidea en las salidas y se trata la normalización necesaria para considerar las salidas de un MLP como probabilidades del HMM (al menos, localmente). Las diferentes ANN son entrenadas por un algoritmo BP estándar pero los errores son calculados por una versión del algoritmo de entrenamiento para HMM modificada para class-HMM. La red para la probabilidad de emisión determina qué probabilidad existe de que el estado en cuestión haya emitido un vector de características dado. La red para las probabilidades de transición para un estado tiene tantas salidas como posibles transiciones posea ese estado. Cada salida determinará la probabilidad de transición hacia ese estado que representa. Por último, la red de etiquetas (o clases) es simplificada en las pruebas ya que se considera sólo una etiqueta posible por estado.

A pesar de que las HNN parecen una simple extensión de los class-HMM, existen varias razones por las que se puede considerar que son potencialmente mejores. En primer lugar, las ANN pueden mapear funciones muy complejas con menos parámetros que las mezclas de gaussianas. Además, las HNN pueden utilizar con más facilidad la información acústica de contexto. Por otro lado, si esta información de contexto también es incorporada en las probabilidades de transición, se aumenta la capacidad para modelar sucesivos segmentos estables conectados por transitorios no-estacionarios. Otro acierto en este enfoque es que, a diferencia de otros trabajos anteriores (por ejemplo (**Renals** y cols., 1994) y (**Robinson**, 1994)), todo el sistema HNN es entrenado de forma conjunta.

Resumen para	(**Krogh** y **Riis**, 1999)
Modelo	Un MLP por estado y por probabilidad a estimar: emisión, transición y clase.
Entradas	13 MFCC, ΔC y varios vectores de contexto.
Salidas	Probabilidades del modelo class-HMM.
Entrenamiento	BP con los errores obtenidos del algoritmo de entrenamiento para HMM
Experimentos	Con diferentes combinaciones de MLP para unas y otras probabilidades. También con diferente cantidad de vectores de contexto.
Base de datos	TIMIT
Reconocimiento	Tipos de fonemas en CSR.
Valoración de resultados	Buenos
Cobertura ASR	60%

Lamentablemente las pruebas sólo incluyen la clasificación de fonemas (en voz continua) en 5 clases de pronunciación (o clases articulatorias). Tampoco se encuentran comparaciones con modelos HMM continuos (solamente con class-HMM discretos con 256 centroides de cuantización vectorial). La base de datos utilizada es la TIMIT y la parametrización es 13 MFCC, con sus 13 ΔC. Para las HNN se prueban diversas variantes del algoritmo de entrenamiento. También se realizan pruebas con diferentes cantidades de nodos ocultos para los MLP y con diferentes cantidades de patrones de contexto en las entradas. Otras pruebas se destinan a comparar el sistema con distintas combinaciones donde se usan o no las redes de probabilidades de emisión o de probabilidades de transición. Los resultados finales llegan al 84.4% siendo de 78.4% las referencias de class-HMM.

IV.2. ANN generadoras de entradas para HMM

Bengio y cols. (1992a) presentan un modelo de MNN para el cálculo de las probabilidades de observación de un HMM para ASR. En este caso hay dos aportes importantes: la estructura modular de varias ANN que se encargan de resolver diferentes subproblemas de clasificación fonológicamente divididos; un algoritmo de optimización global que permite el ajuste fino de las ANN para reducir los errores del sistema de ASR completo y aumentar su poder discriminativo.

El sistema de redes neuronales consta de dos redes en el primer nivel cuyas salidas son analizadas por una tercera red en el segundo nivel. En un tercer nivel se encuentra el sistema basado en HMM que recibe como vectores de observación la salida de la última red neuronal. Una de las redes del

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

primer nivel se encarga de realizar una clasificación de características de relevancia fonética: lugar y forma de la articulación. Las entradas a esta red son las 5 energías de la salida de un filtro pasa banda (Butterworth), la energía total y las seis ΔC. Esta ANN posee cuatro capas conectadas completamente (12-30-15-5) y además conexiones con retardos en el tiempo. La otra ANN del nivel 1 se encarga de clasificar las consonantes plosivas. Posee 16 salidas que indican el lugar, la manera de articulación y el grado de sonoridad. Es alimentada con 74 entradas conteniendo: análisis en frecuencia en escala logarítmica, sus ΔC y la energía total en la banda de 60 a 500Hz con su ΔC. La estructura consta de dos capas ocultas y retardos temporales desde sus salidas hacia la última capa. La ANN del nivel 2 se encarga de combinar todas las salidas de las dos redes del nivel 1 para calcular 8 salidas que serán las observaciones de entrada para el HMM. El modelado mediante HMM se realiza con 14 estados y 28 transiciones (todas hacia delante o en el mismo estado). Las distribuciones son modeladas con mezclas de 5 gaussianas.

El entrenamiento se realiza en tres etapas. En la primera etapa se entrenan las redes neuronales independientemente. En la segunda etapa se utilizan las redes neuronales para alimentar al sistema HMM y obtener sus valores iniciales. En la última etapa se realiza una optimización global de todo el sistema por medio de gradiente descendiente sobre un criterio basado en el modelado temporal de los HMM.

Los resultados finales se presentan a nivel de reconocimiento de clases de fonemas (nasales, fricativos, etc.) en voz continua. Se comparan diferentes combinaciones que van desde las ANN solas hasta el sistema híbrido completo con optimización global. En este último caso se llega hasta un 90% de reconocimiento para la base de datos TIMIT. Una ventaja interesante del método es que, cuando se incorpora un nuevo hablante, permite adaptar el sistema solamente mediante los parámetros de las redes neuronales.

Un complemento de este trabajo se encuentra en (**Bengio** y cols., 1992b), donde con otro título, los mismos autores explican aproximadamente lo mismo pero dando especial atención a las redes modulares para la extracción de características y relegando a un segundo plano el algoritmo para la optimización global del sistema.

Resumen para	(**Bengio** y cols., 1992a)
Modelo	MNN fonológicamente diseñadas y dinámicas (RNN) generan las entradas para un reconocimiento HMM.
Entradas	Preprocesamiento especial para cada módulo
Salidas	Una subred para plosivas, una subred para lugar y manera de articulación y una de integración y compresión en componentes principales.
Entrenamiento	Se entrena cada una por separado y después se realiza un ajuste conjunto con el sistema HMM
Experimentos	Se muestran los resultados para cada red separadamente y reconocimiento de clases de fonemas con el sistema completo.
Base de datos	TIMIT
Reconocimiento	Clases de fonemas, en voz continua
Valoración de resultados	Buenos
Cobertura ASR	50%

Deng y cols. (1994) presentan un trabajo en el que un HMM es utilizado para controlar los cambios en los pesos de una ANN (durante el entrenamiento). En esta propuesta cada sílaba del vocabulario se modela con un HMM de 4 estados, en cada uno de los cuales hay un MLP utilizado como NPM. Estos MLP poseen en la capa oculta una combinación de neuronas sigmoideas y lineales, y en la capa de salida solamente lineales. Cada uno de estos MLP representa un segmento cuasi-estacionario de voz. En la capa de entrada el MLP recibe 7 MFCC y en la salida se encuentra la predicción del próximo vector de características.

El entrenamiento se basa en la combinación de un algoritmo de entrenamiento para HMM (*k*-medias segmental) y BP para los MLP. En una primera etapa se realiza una segmentación (por DP) y en una segunda etapa, a partir de los vectores que quedan asignados a cada estado durante la segmentación, se entrena el MLP como NPM. Las pruebas se realizan en el reconocimiento de seis consonantes (stop-E-set) y una vocal (/*i*/) obtenidas de una base de datos propia. Entre cada sílaba hay una pequeña pausa y todos los registros fueron generados por 6 hablantes masculinos.

Los experimentos consisten en la comparación de 4 sistemas: HMM estandar, MLP-HMM con todas las neuronas ocultas lineales, MLP-HMM con todas las neuronas ocultas no-lineales y MLP-HMM con 2 neuronas ocultas lineales y 3 no lineales. Cada sistema fue entrenado y probado para cada hablante separadamente. Entre los resultados para HMM y MLP-HMM lineal no existe una diferencia

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

importante (todos en torno al 84%) pero con el sistema MLP-HMM no lineal se llega a un 88% y con el combinado a un 92.9%.

Resumen para	(**Deng** y cols., 1994)
Modelo	Un MLP funcionando como NPM, por cada estado del HMM.
Entradas	7 MFCC
Salidas	El MLP predice el próximo vector de características.
Entrenamiento	BP como NPM a partir de una segmentación realizada por DP en el HMM
Experimentos	Un HMM estándar de referencia y en la capa intermedia del MLP se usan diferentes combinaciones de funciones de activación lineales y no lineales.
Base de datos	Propia, 6 hablantes masculinos
Reconocimiento	Fonema /*i*/ y stop-E-set, dependiente del hablante.
Valoración de resultados	Buenos
Cobertura ASR	60%

Siguiendo en la línea de las ANN para asistir a los HMM, **Chung** y **Un** (1996a) presentan otro modelo híbrido, donde varios MLP transforman los vectores de características que alimentan al HMM. En este modelo cada MLP utiliza información de contexto para realizar una predicción del próximo vector de características de voz y el HMM es alimentado por el error de predicción. De esta forma el MLP es utilizado como NPM. La elección de los autores de asignar un MLP para cada estado del HMM está de acuerdo con el hecho de que un único MLP modela mal la dinámica no estacionaria de la señal de voz.

Para el entrenamiento es necesario un primer paso de alineación de los modelos mediante DP. De esta forma no se requiere ninguna información de antemano acerca de la posición de los fonemas. Después de la alineación los MLP para cada estado son entrenados mediante un algoritmo de BP (la salida es conocida porque en el NPM se trata del mismo vector de características de la voz para el tiempo dado). Estos dos pasos, junto con la optimización de las probabilidades del HMM, se iteran hasta satisfacer cierto criterio de convergencia. Para ampliar la capacidad discriminativa del modelo híbrido los autores formulan una función objetivo que contempla todo el error de predicción a lo largo de la mejor secuencia para cada frase de entrenamiento. Con esta modificación se pueden adaptar los pesos de cada MLP por BP mediante el gradiente de la función objetivo con respecto a todos los errores de predicción de la secuencia.

Las pruebas se realizan sobre una base de datos de 102 palabras coreanas: días, meses y horarios. La base de datos fue generada por un total de 26 hablantes y como parametrización se utilizaron CC, energía y sus ΔC. Es importante notar que al comparar el sistema con otro reconocedor estándar basado en HMM (con la misma estructura), si bien los errores durante el entrenamiento son menores, cuando se realizan las pruebas de validación el sistema HMM estándar es el mejor.

Resumen para	(**Chung** y **Un**, 1996a)
Modelo	Un MLP por estado del HMM se entrena como NPM y se alimenta el HMM con el error de predicción.
Entradas	CC, energía, sus ΔC con varios vectores de contexto.
Salidas	Predicción del próximo vector de características.
Entrenamiento	MLP por BP y un entrenamiento discriminativo del sistema completo
Experimentos	Con y sin entrenamiento discriminativo y en relación a un sistema HMM estándar de referencia.
Base de datos	Propia, en coreano, contiene meses, días y horarios.
Reconocimiento	CSR basado en HMM.
Valoración de resultados	Regulares
Cobertura ASR	80%

En el trabajo de **Jang** y **Un** (1996) se presenta un modelo híbrido TDNN-HMM donde las salidas de la capa media de la TDNN se utilizan como un cuantizador vectorial difuso (FVQ), del que se alimentan los HMM y se obtiene al mismo tiempo un suavizado de las probabilidades de observación.

Cuando en un HMM discreto no se tiene la suficiente cantidad de datos para el entrenamiento es necesario realizar un proceso de suavizado de las probabilidades de observación de los diferentes símbolos. En el método de suavizado por umbral, cuando una probabilidad es menor que cierto valor se trunca a este valor mínimo. Este método es de muy simple implementación, pero tiene un efecto de suavizado muy pobre. Existen otros métodos estadísticos más elaborados para realizar este suavizado, sin embargo, los autores proponen extraer una matriz de suavizado de los valores de activación de los nodos de la capa oculta de una TDNN entrenada previa e independientemente.

La TDNN posee una estructura modular en cuya capa oculta se encuentran subredes para clasificar 11 fonemas y subredes para clasificar en consonante o vocal. Esta estructura modular es duplicada

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

de forma que cada conjunto de redes se corresponde con un estado del HMM y se aboca a intervalos consecutivos de la señal de entrada. Cada subred es entrenada independientemente y luego las conexiones entre las subredes y las restantes capas de la TDNN se entrenan (BPTT) de forma de obtener un vector de salidas para cada estado del HMM. De esta forma, la segunda capa de múltiples subredes se comporta como un sistema de cuantización vectorial suavizado o difuso (al menos en el sentido que le dan los autores a este término). Las ecuaciones de entrenamiento y evaluación de los HMM son reformuladas para manejar estos nuevos vectores.

Resumen para	(**Jang** y **Un**, 1996)
Modelo	La segunda capa de una TDNN genera los vectores de observación del HMM y provee de una matriz de suavizado.
Entradas	20 CC con contexto.
Salidas	Clasificación en fonemas y vocal/consonante.
Entrenamiento	BPTT y propio de HMM
Experimentos	Sintonización preliminar y reconocimiento con y sin suavizado.
Base de datos	Propia, en coreano, pocos hablantes.
Reconocimiento	Palabras aisladas, basado en HMM.
Valoración de resultados	Buenos
Cobertura ASR	50%

Para las pruebas se utiliza una base de datos en coreano que contiene 75 palabras aisladas pronunciadas por 7 hablantes masculinos. Los vectores de características se extraen cada 10 ms para obtener 20 CC. La segmentación para el entrenamiento se hace manualmente. Los modelos de palabras se construyen concatenando los HMM de fonemas. En los resultados finales se compara el modelo con suavizado por umbral y el modelo con el suavizado propuesto (también se presentan otros resultados de sintonización preliminares). Se parte de un error muy elevado (44%) que posiblemente sea debido a la mala estimación de las probabilidades de observación para los HMM (por insuficientes datos de entrenamiento). Los resultados con el método de suavizado propuesto reducen el error a 10.7%.

IV.3. Otras combinaciones ANN-HMM

Chung y **Un** (1996b) presentan otra alternativa de combinación de MLP y HMM. El MLP es entrenado para la clasificación de fonemas y luego se utilizan sus salidas para realizar un pesado de las gaussianas que se mezclan para modelar las distribuciones de probabilidad de observación en cada estado de un HMM estándar. Para aumentar el poder discriminativo se propone un algoritmo de optimización global basado en gradiente descendiente.

En este trabajo, a diferencia de los trabajos de Bourlard y cols., el MLP provee el valor de un exponente adicional al peso de las gaussianas que se mezclan para obtener la probabilidad de observación. Este exponente es función del vector de entrada y del estado del modelo HMM del que se trata. Para incorporar más información se contexto se utilizan dos MLP con 28 neuronas de salida, una para cada fonema. El exponente de pesado se obtiene multiplicando las salidas de los dos MLP, considerando el efecto que tendrían sobre una base mayor o menor que 1. En primer lugar, se entrenan los MLP separadamente del HMM y a continuación se define una función de costo para el entrenamiento discriminativo que luego es optimizada por BP en los MLP.

Resumen para	(**Chung** y **Un**, 1996b)
Modelo	2 MLP entrenados para clasificación de fonemas se usan para pesar las gaussianas de las probabilidades de observación de un HMM.
Entradas	CC, energía y sus ΔC.
Salidas	28 fonemas en los MLP.
Entrenamiento	MLP por BP y un entrenamiento discriminativo del sistema completo
Experimentos	MLP para clasificación de fonemas, sistema integrado en habla continua.
Base de datos	Propia, en coreano, contiene meses, días y horarios.
Reconocimiento	CSR basado en HMM.
Valoración de resultados	Regulares
Cobertura ASR	80%

El sistema completo es probado con una base de datos de 102 palabras coreanas en un contexto bastante reducido: días, meses y horarios (ya citada anteriormente). En total se utilizan 16 hablantes para el conjunto de entrenamiento y otros 10 para el de prueba. Los vectores de características poseen 26 elementos entre los que se encuentran CC, energía y sus ΔC. Entre los resultados se muestra el desempeño de los MLP para clasificación de fonemas independientemente del sistema HMM. Luego, con el sistema completo se logra una reducción porcentual del error de reconocimiento

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

en palabras del 30% (de 15.9 a 11.1%). Todo esto a pesar de que la clasificación del MLP apenas supera el 60%.

Kurimo (1997) (desde la Universidad de Helsinki) propone la utilización de SOM y LVQ para el entrenamiento de las mezclas de densidades de probabilidad de observación en sistemas HMM. Por un lado se utilizan los SOM para producir una agrupación de los vectores de características para cada estado de un HMM y luego para la inicialización de los vectores de medias de su mezclas de gaussianas. Para esto se utiliza un gran SOM que cubre todos los fonemas y luego se divide éste en varios conjuntos de centroides de cuantización vectorial separados para cada fonema de acuerdo a la etiqueta que posean.

Por otro lado se utiliza el algoritmo LVQ (en particular LVQ3) para entrenar el conjunto de gaussianas de forma de maximizar la discriminación en los bordes de las áreas correspondientes a los distintos fonemas (reemplazando al comúnmente utilizado *k*-medias).

Para las pruebas se utilizan dos bases de datos en finlandés, una para entrenamiento, con tres hablantes y otra para las pruebas, con 7 hablantes. En todos los casos se trata de reconocimiento de palabras aisladas con una parametrización MFCC de 20 coeficientes con diferentes configuraciones de contexto. Los resultados se presentan para un amplio espectro de algoritmos y son adecuadamente analizados para determinar la significancia estadística de las mejoras encontradas.

Resumen para	(**Kurimo**, 1997)
Modelo	SOM y LVQ para la inicialización y entrenamiento de distribuciones de probabilidad de los HMM.
Entradas	MFCC, con contexto y sus ΔC.
Salidas	Los sistemas neuronales tienen distintas funciones en el HMM, no necesariamente una clasificación.
Entrenamiento	Modificación de los métodos clásicos de HMM para incorporar SOM y LVQ3.
Experimentos	Muy completos, incluyen preprocesamiento, inicialización y entrenamiento.
Base de datos	Propia, en finlandés.
Reconocimiento	Fonemas, basado en HMM.
Valoración de resultados	Muy buenos
Cobertura ASR	50%

En cuanto a los métodos de inicialización se pudo reducir considerablemente la cantidad de iteraciones y el error final mediante la aplicación de una combinación de SOM y LVQ3. En cuanto a los métodos de entrenamiento, para la misma cantidad de reestimaciones, se obtienen mejoras significativas. También se incluyen resultados de diferentes pruebas donde se varía el preprocesamiento, incluyendo diferentes combinaciones de vectores de contexto y ΔC. En el mejor de los casos se obtiene una reducción en el error de hasta el 40%.

Si bien el área de reconocimiento robusto del habla (RSR) está comenzando a considerarse como un área nueva y separada del ASR, se consideró oportuno incorporar el trabajo de **Moon** y **Hwang** (1997) en este artículo, debido a que el modelo propuesto está en relación directa con las combinaciones ANN-ASR. Los métodos de compensación para RSR pueden ser clasificados en cuatro grandes grupos. En el primer grupo están las técnicas de filtrado y realce de la voz. El segundo grupo intenta ajustar los parámetros de los reconocedores de forma que los modelos se adapten a las condiciones de ruido. Una tercera categoría está orientada a reducir la sensibilidad del sistema de reconocimiento ante posibles ruidos en las entradas. En la última categoría se incorporan las estadísticas del ruido en los parámetros del modelo en la misma fase de entrenamiento.

En este artículo los autores proponen un algoritmo de inversión de los HMM que ha sido inspirado en la inversión de ANN, viendo a los HMM como un caso especial de ANN. En la inversión de una ANN se encuentran las entradas que optimizan algún criterio mientras se mantienen fijos los pesos. La inversión de los HMM se aplica al RSR moviendo las entradas de voz hacia las medias de las mezclas de gaussianas (con las apropiadas restricciones).

Las pruebas están orientadas al RSR para diferentes condiciones de contaminación de la señal de voz. En particular se utiliza una base de datos de dígitos aislados. Los resultados están contrastados con otros métodos y si bien las mejoras no son muy significativas, el ahorro en tiempo de cómputo es muy importante en relación a métodos con similar desempeño.

Chen y cols. (1998) utilizan una RNN con una máquina de estados finitos (FSM) para reducir el espacio de búsqueda en un reconocedor basado en HMM. Se trata de otra alternativa en la combinación ANN-HMM, en este caso no existe una integración tan afín de los dos paradigmas ya que el entrenamiento es independiente.

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

Este sistema está orientado al reconocimiento de sílabas del mandarín. El módulo de preclasificación detecta ciertas peculiaridades fonológicas de las sílabas que son importantes en este lenguaje (y otros lenguajes tonales de oriente). El módulo RNN-FSM se encarga de clasificar y segmentar la entrada de voz según cuatro clases: inicio de sílaba, fin de sílaba, silencio y estado transitorio. La RNN posee una capa oculta con 25 neuronas y realimentación con un retardo en el tiempo desde esta capa hacia la entrada. Por lo tanto, la entrada queda compuesta por las características de la voz y la retroalimentación desde la capa oculta. La capa de salida posee 3 neuronas (silencio e inicio y fin de sílaba). Este clasificador neuronal trabaja al nivel (temporal) de un segmento de adquisición y es entrenado con una extensión del algoritmo de BP. Luego la FSM se encarga de la secuencia de clasificación y detección de posibles estados transitorios. Esta FSM se basa en la evaluación de varios umbrales (determinados empíricamente) de aceptación de las salidas de la RNN.

Finalmente, se utiliza un reconocedor HMM estándar pero se restringe el espacio de búsqueda de acuerdo a la secuencia de preclasificación. Es decir, en lugar de buscar, por ejemplo, la probabilidad más alta entre todos los fonemas posibles, si la preclasificación determina para un tiempo dado que se trata de un inicio de sílaba, entonces se buscará solamente entre los fonemas que pueden ser inicio de sílaba (conocimiento lingüístico que se posee a priori).

Resumen para	(**Chen** y cols., 1998)
Modelo	RNN pequeña asistida por una FSM para el procesamiento de la secuencia y reducción del espacio de búsqueda en un sistema de ASR-HMM
Entradas	MFCC, energía y sus ΔC.
Salidas	Silencio, inicio o fin de sílaba y estado transitorio.
Entrenamiento	Extensión de BP para la RNN y umbrales fijados empíricamente para la FSM. Los HMM por separado.
Experimentos	RNN sola, con FSM y sistema RNN-FSM-HMM completo para reconocimiento de sílabas del mandarín.
Base de datos	Propia, un hablante masculino
Reconocimiento	Sílabas, voz continua.
Valoración de resultados	Regulares
Cobertura ASR	30%

El sistema de reconocimiento HMM posee 411 modelos para las diferentes sílabas. Los modelos consisten en 8 estados que resultan de la concatenación de 3 estados para los fonemas de inicio de sílaba y 5 estados para los de fin de sílaba. Además existe un modelo de un estado para el silencio. Para cada estado se utilizan desde 1 a 8 gaussianas dependiendo de la cantidad de datos disponibles para el entrenamiento.

La base de datos es propia y de un sólo hablante masculino. Se realiza un preprocesamiento estándar basado en MFCC, energía y sus ΔC. Los resultados no son mejores que la referencia (el sistema basado en HMM puro) pero los autores apuntan a los beneficios en el tiempo de cómputo al restringir el espacio de búsqueda mediante la preclasificación.

V. Discusión y Conclusiones

No son pocos los aspectos que hay para discutir. De forma individual y algunas veces comparativa, se ha incluido en la descripción de cada artículo una discusión en torno a los enfoques propuestos, el desarrollo y los resultados. También en forma individual se presentan el resumen y las conclusiones al final de cada artículo, con una extensión que depende de su importancia. En esta sección se tratará de abordar la discusión desde una óptica más amplia, tratando de destacar las características más importantes de este desarrollo.

V.1. El problema del ASR

En este apartado se pretende, brevemente, establecer un marco conceptual y terminológico para las siguientes discusiones. Para esto se propone dividir el problema del ASR en dos grandes subproblemas que denominaremos: *discriminación* y *dinámica* (DD). Es cierto que estos dos problemas no se presentan totalmente separados y que en ocasiones es difícil distinguir donde termina uno y comienza el otro. Sin embargo será útil mantener esta distinción conceptual.

El problema de la discriminación consiste en encontrar un sistema capaz de distinguir entre unos y otros segmentos característicos del habla. No es necesario que estos segmentos se correspondan con alguna estructura lingüística conocida, como el fonema o la palabra, pero en los sistemas de reconocimiento hay cierta tendencia a que el modelo involucre estas unidades. Tampoco es

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

necesario que esta discriminación se vea separadamente de la dinámica, porque también la dinámica aporta información útil para la discriminación.
En cuanto a la dinámica, se trata de modelar o capturar las variaciones de las características del habla en el tiempo. Nuevamente, no se pueden separar por completo los problemas DD porque cuanto más diferenciadas sean esas características, más fácil será seguir su evolución en el tiempo.
Sin embargo, en cuanto a la dinámica, muchas veces se han seguido caminos totalmente separados de la discriminación, por ejemplo, en el modelado de secuencias (discretas) en el tiempo.
Los problemas DD deben ser considerados a diferentes *escalas de observación* y no simplemente a nivel de vectores de características como muchos proponen. Partiendo de los vectores de características se puede ascender hacia unidades subfonémicas, fonemas, suprasegmentos, sílabas, morfemas, palabras, funciones de palabras, frases subordinadas, frases y quizás sería posible llegar a niveles de abstracción mucho más altos, entrando por ejemplo en el terreno de los sistemas de diálogo, la semántica y pragmática, las emociones y más.
Pero estas divisiones no tienen porque formar compartimentos estancos, y es necesario *modelar la interacción* entre los diferentes niveles y las escalas de observación. Como un ejemplo sencillo, actualmente se modelan como distintos fonemas aquellos que se encuentran en distintos contextos de fonemas o inmersos en diferentes palabras.

V.2. Tipos anómalos

Con estas críticas de este apartado no se pretende caer en la situación de los críticos del relato citado en por **Bourlard** y cols. (1996): “Cuando un nuevo descubrimiento se notifica, lo que primero dicen los críticos es «debe ser falso». Cuando pasado el tiempo se demuestra que es totalmente verdaderos, los críticos dicen «será verdadero pero no es muy importante». Y cuando después de muchos años los desarrollos y las aplicaciones dejan clara su importancia, los críticos dirán: «será importante, pero ya no es demasiado novedoso».” Sin embargo, hay que destacar algunas peculiaridades negativas en los enfoques de ciertos investigadores. Hay trabajos en que se emplea la denominación de “redes neuronales” sólo para usar su pomposidad y encuadrarlos en un área de gran crecimiento en la década. Los modelos que proponen, muchas veces tienen poco de “neuronales” aunque si son “redes” o modelos conexionistas (que es un concepto más amplio). En ocasiones se encuentran marcados desequilibrios. Por ejemplo hay trabajos en los que se profundizan muchos aspectos fonológicos para el sistema de preprocesamiento y luego se desperdician en un sistema muy elemental de reconocimiento. Y otros que, en el extremo opuesto, aplican por “fuerza bruta” un sistema inteligente a una señal que no contiene la información que se pretende extraer, esperando que el sistema inteligente “haga magia”. Hay que destacar que en algunos trabajos los resultados no se obtienen con un método suficientemente fiable, o bien no se hace mención acerca de la significancia estadística de las diferencias encontradas con otros paradigmas. Hay trabajos muy buenos pero que no cubren mucho del ASR, hasta trabajos que parecen cubrir muchos aspectos del ASR pero tienen varios puntos oscuros. Hay trabajos muy primitivos, cuando un MLP era “toda una novedad”, hasta trabajos con sistemas muy complejos, especialmente diseñados para el ASR y con muy buenos resultados. De todos los casos comentados en este apartado se puede encontrar al menos un ejemplo en el presente artículo.

V.3. Los dos paradigmas por separado

Un gran acierto en el paradigma de los HMM es que varias de las escalas mencionadas anteriormente son incorporadas bajo la misma filosofía de autómatas probabilísticos en el proceso de programación dinámica. De esta, con mayor o menor éxito, los problemas de DD se tienen en cuenta conjuntamente. Entre las mayores deficiencias de los HMM en relación a las ANN se pueden citar: su baja capacidad en la discriminación estática, la hipótesis de que la secuencia de estados es una cadena de Markov de primer orden, la incorporación de información de contexto en los vectores de características aumenta demasiado (cuadráticamente) la complejidad del sistema y, finalmente, en los HMM se deben resolver de forma empírica muchas cuestiones relativas a la topología (lo cual suele ser una ventaja en algunos casos). En cuanto a las ANN, además de las ventajas comparativas que se desprenden de lo anterior, hay que considerar que no poseen tantas restricciones acerca de las distribuciones estadísticas de los patrones de entrenamiento y poseen una mayor capacidad de generalización. En contraste, las ANN siguen siendo mucho más ineficientes a la hora de modelar las características dinámicas del habla. A pesar de que las RNN pueden resolver el problema de modelado temporal sin necesidad de recurrir al costoso proceso de programación dinámica, en general esto sólo es útil a nivel local y no permite generalizarse hacia

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

otras escalas de observación. Cuando se consideran las ANN puras, difícilmente se logra una integración de las escalas de observación con una amplitud semejante a la de los HMM. Difícilmente una palabra que se encuentra un segundo más adelante modifique la decisión acerca de un fonema que se está evaluando en el instante actual. En todo caso, los modelos neuronales puros han llegado a alcanzar la mitad de la cobertura del problema de ASR. Existen raras excepciones que se presentan como modelos neuronales pero más bien son versiones alternativas de HMM y siempre dependientes de un algoritmo "no neuronal" de programación dinámica.
Hay muchos trabajos que muestran muy buenos resultados en la clasificación de fonemas olvidándose por completo del problema de la dinámica a escalas superiores, y esto relativiza totalmente los buenos resultados porque el comportamiento de esos sistemas basados en una mala segmentación sería realmente impredecible. Sin embargo, los defensores de las ANN muchas veces insisten en el siguiente argumento: "El mejor sistema de ASR que existe actualmente es neuronal" (el cerebro humano). Esto da pie para pensar que el camino correcto hacia la solución del problema de ASR está en los sistemas de ANN. Sin embargo hay un contraejemplo que debilita esta postura: "¿Acaso el mejor sistema para volar que existe actualmente mueve las alas?" Es de esperar que en ninguno de los dos extremos esté la solución definitiva.
Finalmente, en relación con las ANN, se ha observado una creciente utilización de RNN (hasta 1995) y una acertada incorporación de sistemas MNN.

V.4. Los dos paradigmas juntos

Ya que por separado los paradigmas tienen ventajas y desventajas en muchos casos disjuntas, unirlos resultó ser una buena alternativa. Los métodos propuestos para utilizar conjuntamente los dos paradigmas son muy variados. Aunque parezca un juego de palabras y a veces sea difícil ubicar el ejemplo concreto de cada caso, se enumerarán a continuación los distintos métodos de convergencia de paradigmas. En primer lugar están los que muestran a un paradigma como si fuera el otro o bien, como una generalización o inclusión del otro. En otro grupo están las ideas que se dirigen a reemplazar partes de un paradigma con elementos sencillos del otro. También, un poco menos pretenciosos, están los enfoques que usan las salidas de un sistema para alimentar al otro. Y finalmente, están los sistemas totalmente integrados, en teoría e implementación, donde cuesta distinguir donde termina un paradigma y comienza el otro. Quizás, uno de los aspectos más interesantes de estas combinaciones es que los sistemas que son basados en HMM, aunque se prueben en palabras aisladas o fonemas, son directamente generalizables hacia el reconocimiento del habla continua.
En general, mediante una representación del espacio de estados (aplicable tanto a los HMM como a las RNN) se puede dividir el funcionamiento de la integración en dos pasos principales: la actualización de las variables de estado y el cálculo o predicción de la salida dada la secuencia de estados. Evidentemente se sigue tratando de manejar procesos de naturaleza discreta, que guardan mucha información contextual a lo largo de una cantidad (teóricamente) indefinida de tiempo.
Para terminar planteamos algunas preguntas: ¿No estamos simplemente en el círculo vicioso de la constante sintonización de los HMM? ¿No sería mejor, en vez de tratar de imitar o sintonizar los HMM, buscar nuevas estrategias? Como las arquitecturas raramente son inferidas por los datos: ¿Para el modelado en el tiempo deberían considerarse arquitecturas que se adapten a las características dinámicas de los datos?

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

Apéndice: acrónimos

Acrónimo	*Significado en Inglés*	*Significado en Español*
ASR	Automatic Speech Recognition	Reconocimiento automático del habla
CSR	Continuous Speech Recognition	Reconocimiento del habla continua
RASR	Robust Automatic Speech Recognition	Reconocimiento automático del habla robusto
HMM	Hidden Markov Models	Modelos ocultos de Markov
DHMM	Discrete Hidden Markov Models	Modelos ocultos de Markov discretos
CHMM	Continuous Hidden Markov Models	Modelos ocultos de Markov continuos
ANN	Artificial Neural Networks	Redes neuronales artificiales
NPM	Neural Predictive Model	Modelo neuronal predictivo
MNN	Modular Neural Networks	Redes neuronales (artificiales) modulares
RNN	Recurrent Neural Networks	Redes neuronales recurrentes
PNN	Probabilistic Neural Network	Red neuronal probabilística

PRNN	Partially Recurrent Neural Networks	Redes neuronales recurrentes parciales
CRNN	Chaotic Recurrent Neural Networks	Redes neuronales recurrentes y caóticas
PCMN	Partially Connected Multilayer Network	Red multicapa parcialmente conectada
TDNN	Time-Delay Neural Networks	Redes neuronales con retardo en el tiempo
TLTDNN	Two-Level Time-Delay Neural Networks	Redes neuronales con retardo en el tiempo de dos niveles
SOM	Self-Organizing Maps	Mapas autoorganizativos
LVQ	Learning Vector Quantization	Cuantización vectorial con aprendizaje
MLP	Multi-Layer Perceptrons	Perceptrones multicapa
RBFN	Radial Basis Function Networks	Redes con funciones de base radial
TDRBFN	Time Delay Radial Basis Function Networks	Redes con funciones de base radial y retardos temporales
ODWE	Orthogonal Delta Weight Estimator	Estimador ortogonal de variación de pesos
HCNN	Hidden Control Neural Networks	Redes neuronales con control oculto
HNN	Hidden Neural Networks	Redes neuronales ocultas
IOHMM	Input-Output Hidden Markov Models	Modelos de Markov de entrada-salida
IJNN	Intelligent Judge Neural Networks	Redes neuronales de juicio inteligente
LMS	Least Mean Square	Mínimos cuadrados medios
BP	Back-Propagation	Retro-propagación
BPTT	Back-Propagation Through Time	Retro-propagación a través del tiempo
GRBF	Generalized Radial Basis Function Networks	Redes con funciones de base radial generalizadas
DCL	Differential competitive learning	Aprendizaje competitivo diferencial
DP	Dynamic Programming	Programación dinámica
FSM	Finite State Machine	Máquina de estados finitos
SC	Spectral Coefficients	Coeficientes espectrales
MSC	mel-scale Spectral Coefficients	Coeficientes espectrales en escala de mel
STFT	Short-Time Fourier Transform	Transformada de Fourier de tiempo corto
CC	Cepstral Coefficients	Coeficientes cepstrales
MFCC	Mel Frequency Cepstral Coefficients	Coeficientes cepstrales en escala de mel
ΔC	Delta Coefficients	Coeficientes delta (primera derivada en el tiempo)

Notas y referencias

[1] Para facilitar la lectura del texto y el acceso a la bibliografía citada se han utilizado las siglas originales en inglés. En el Apéndice se puede encontrar una tabla con todos los acrónimos utilizados.

[2] No se incluye el ASR en condiciones adversas (reconocimiento robusto) en esta valoración.

[3] En las tablas estos títulos se han abreviado como: Entradas (y preprocesamiento), Salidas (o clasificación), Reconocimiento (tipo de ASR), Valoración de resultados (valoración global de los resultados), Cobertura ASR (cobertura del problema de ASR).

[4] Nodos en la capa de entrada, la oculta y la de salida respectivamente.

[5] Fenómeno poco frecuente, aunque entendible en este caso dado que se utiliza una población muy pequeña

[6] Esto no está claro en el artículo original (ya que es muy corto), pero como el nivel superior recibe las salidas del inferior –clasificación y predicción– se puede suponer que se entrena con la clasificación correcta en lugar de la predicción de la entrada. Sobre el entrenamiento o la definición de reglas del sistema difuso no se hace ninguna mención.

[7] Disponible en: http://htk.eng.cam.ac.uk

[8] Se conservará la abreviatura class-HMM en lugar de la utilizada por los autores (CHMM), para evitar confusiones con la utilización más frecuente de esta sigla para Countinuous Hidden Markov Models.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

Bibliografía

BALDI P. y CHAUVIN Y. “Hybrid modeling HMM/NN architectures and protein applications”. **Neural Computation**, 8(7)1996:1541-1565.

BANBROOK M., MCLAUGHLIN S., y MANN I. “Speech characterization and synthesis by nonlinear methods”. **IEEE Transactions on Speech and Audio Processing**, 7(1)1999:1-17.

BENGIO Y. y FRASCONI P. “Input-output HMM´s for sequence processing”. **IEEE Transactions on Neural Networks**, 7(5)1996:1231-1249.

BENGIO Y., MORI R. D., FLAMMIA G., y KOMPE R. “Global optimization of a neural network-hidden Markov model hybrid”. **IEEE Transactions on Neural Networks**, 3(2)1992a:252-259.

BENGIO Y., MORI R. D., FLAMMIA G., y KOMPE R. “Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks”. **Speech Communication**, 11(2)1992b:261-271.

BOURLARD H., HERMANSKY H., y MORGAN N. “Towards increasing speech recognition error rates”. **Speech Communication**, 18(3)1996:205-231.

BOURLARD H. y MORGAN N. “Continuous speech recognition by connectionist statistical methods”. **IEEE Transactions on Neural Networks**, 4(6)1993:893-909.

BOURLARD H., MORGAN N., y RENALS S. “Neural nets and hidden Markov models: Review and generalizations”. **Speech Communication**, 11(2)1992:237-246.

BOURLARD H. y WELLEKENS C. J. “Speech pattern discrimination and multilayer perceptrons”. **Computer Speech and Language**, 3(1)1989:1-19.

BOURLARD H. y WELLEKENS C. J. “Links betwen Markov models and multilayer perceptrons”. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 12(12)1990:1167-1178.

BRIDLE J. S. “Alpha-nets: a recurrent ` neural' network architecture with a hidden Markov model interpretation”. **Speech Communication**, 9(1)1990:83-92.

CECCARELLI M. y HOUNSOU J. T. “Sequence recognition with radial basis function networks: experiments with spoken digits”. **Neurocomputing**, 11(1)1996:75-88.

CHEN S.-H. y LIAO Y.-F. “Modular recurrent neural networks for Mandarin syllable recognition”. **IEEE Transactions on Neural Networks**, 9(6)1998:1430-1441.

CHEN S.-H., LIAO Y.-F., CHIANG S.-M., y CHANG S. “An RNN-based preclassification method for fast continuous mandarin speech recognition”. **IEEE Transactions on Speech and Audio Processing**, 6(1)1998:86-90.

CHUDÝ V., HAPÁK V., y CHUDÝ L. “Isolated word recognition in Slovak via neural nets”. **Neurocomputing**, 3(5)1991:259-282.

CHUNG Y. J. y UN C. K. “An MLP/HMM hybrid model using nonlinear predictors”. **Speech Communication**, 19(4)1996a:307-316.

CHUNG Y. J. y UN C. K. “Multilayer perceptrons for state-dependent weightings of HMM likelihoods”. **Speech Communication**, 18(1)1996b:79-89.

COLE R. A., NOEL M., LANDER T., y DURHAM T. “New telephone speech corpora at CSLU”. En **Proceedings of 4th European Conference of Speech Communication and Technology**. 1995:821-824.

COSI P., BENGIO Y., y DEMORI R. “Phonetically-based multi-layered neural networks for vowel classification”. **Speech Communication**, 9(1)1990:15-29.

COSI P., DUGATTO M., FERRERO F., CALDOGNETTO E. M., y VAGGES K. “Phonetic recognition by recurrent neural networks working on audio and visual information”. **Speech Communication**, 19(3)1996:245-252.

DELLER J., PROAKIS J., y HANSEN J. **Discrete Time Processing of Speech Signals**. Macmillan Publishing, NewYork. 1993.

DENG L., HASSANEIN K., y ELMASRY M. “Analysis of the correlation structure for a neural predictive model with application to speech recognition”. **Neural Networks**, 7(2)1994:331-339.

DJEZZAR L. y PICAN N. “Phonetic knowledge embedded in a context sensitive MLP for french speaker-independent speech recognition”. **Speech Communication**, 21(3)1997:155-167.

ELMAN J. L. “Finding structure in time”. Technical Report 8801, CRL. 1988.

FREEMAN J. y SKAPURA D. **Neural Networks. Algorithms Aplications and Programming Techniques**. Addison-Wesley Publishing Company.

GRAMSS T. y STRUBE H. W. “Recognition of isolated words based on psychoacoustics and neurobiology”. **Speech Communication**, 9(1)1990:35-40.

GREENWOOD G. W. “Training partially recurrent neural networks using evolutionary strategies”. **IEEE Transactions on Speech and Audio Processing**, 5(2)1997:192-194.

HANES M. D., AHALT S., y KRISHNAMURTHY A. “Acoustic-to-phonetic mapping using recurrent neural networks”. **IEEE Transactions on Neural Networks**, 5(4)1994:659-662.

HAYKIN S. **Neural Networks. A Comprehensive Foundation**. Macmillan College Publishing Company.

IRINO T. y KAWAHARA H. “A method for designing neural networks using nonlinear multivariate analysis: Application to speaker-independent vowel recognition”. **Neural Computation**, 2(3)1990:386-397.

JANG C. S. y UN C. K. “A new parameter smoothing method in the hybrid TDNN/HMM architecture for speech recognition”. **Speech Communication**, 19(4)1996:317-324.

JELINEK F. **Statistical Methods for Speech Recognition**. MIT Press.

JORDAN M. L. “Serial order: A parallel distributed processing approach”. Technical Report 8604, CRL.

KIM D.-S. y LEE S.-Y. “Intelligent judge neural network for speech recognition”. **Neural Processing Letters**, 1(1)1994:17-20.

sinc(*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

KNAGENHJELM P. y BRAUER P. “Classification of vowels in continuous speech using MLP and a hybrid net”. **Speech Communication**, 9(1)1990:31-34.
KOHONEN T. “The self-organizing map”. **Proceedings of the IEEE**, 78(9)1990:1464-1480.
KOHONEN T. **The Self-Organizing Map**. Springer-Verlag.
KOHONEN T., MAKISARA K., y SARAMAKI T. “Phonotopics maps - insightful representation of phonological features for speech recognition”. En **IEEE Proceedings of the 7th International Conference on Pattern Recognition**. 1984:182-185.
KONG S.-G. y KOSKO B. “Differential competitive learning for centroid estimation and phoneme recognition”. **IEEE Transactions on Neural Networks**, 2(1)1991:118-124.
KROGH A. y RIIS S. K. “Hidden Neural Networks”. **Neural Computation**, 11(2)1999:541-563.
KUHN G., WATROUS R. L., y LADENDORF B. “Connected recognition with a recurrent network”. **Speech Communication**, 9(1)1990:41-48.
KURIMO M. “Training mixture density HMMs with SOM and LVQ”. **Computer Speech and Language**, 11(4)1997:321-343.
LEE T. y CHING P. C. “Cantonese syllable recognition using neural networks”. **IEEE Transactions on Speech and Audio Processing**, 7(4)1999:466-472.
LEVIN E. “Hidden control neural architecture modeling of nonlinear time varying systems and its applications”. **IEEE Transactions on Neural Networks**, 4(1)1993:109-116.
LIPPMANN R. P. “An introduction to computing with neural nets”. **IEEE Acoustics Speech and Signal Processing Magazine.**, 4(2)1987:4-22.
LIPPMANN R. P. “Review of neural networks for speech recognition”. **Neural Computation**, 1(1)1989:1-38.
LIPPMANN R. P. “Speech recognition by machines and humans”. **Speech Communication**, 22(1)1997:1-15.
MARTENS J. y DEPUYDT L. “Broad phonetic classification and segmentation of continuous speech by means of neural networks and dynamic programming”. **Speech Communication**, 10(1)1991:81-90.
MOON S. y HWANG J.-N. “Robust speech recognition based on joint model and feature space optimization of hidden Markov models”. **IEEE Transactions on Neural Networks**, 8(2)1997:194-204.
NGUYEN M. H. y COTTRELL G. W. “Tau net. a neural network for modeling temporal variability”. **Neurocomputing**, 15(3)1997:249-271.
PETEK B., WAIBEL A., y TEBELSKIS J. M. “Integrated phoneme and function word architecture of hidden control neural networks for continuous speech recognition”. **Speech Communication**, 11(2)1992:273-282.
PETERSON G. E. y BARNEY H. L. “Control methods used in a study of the vowels”. **J. Acoustical Society of America**, 24(2)1952:175-184.
PITRELLI J. “Phonebook: A phonetically-rich isolated-word telephone-speech database”. En **Proceedings of International Conference of Acoustics, Speech and Signal Processing**. 1995:101-104.
POO G.-S. “Large vocabulary mandarin final recognition based on two-level time-delay neural networks (TLTDNN)”. **Speech Communication**, 22(1)1997:17-24.
PRICE P., FISHER W. M., BERNSTEIN J., y PALLET D. S. “The DARPA 1000-word resource management database for continuous speech recognition”. En **Proceedings of International Conference of Acoustics Speech and Signal Processing**. 1988:651-654.
RABINER L. R. y GOLD B. **Theory and Application of Digital Signal Processing**. Prentice Hall.
RABINER L. R. y JUANG B. H. “An introduction to hidden Markov models”. **IEEE Acoustics Speech and Signal Processing Magazine**, 3(1)1986:4-16.
RABINER L. R. y JUANG B. H. **Fundamentals of Speech Recognition**. Prentice Hall.
RENALS S., MORGAN N., BOURLARD H., COHEN M., y FRANCO H. “Connectionist probability estimators in HMM speech recognition”. **IEEE Transactions on Speech and Audio Processing**, 2(1)1994:161-174.
ROBINSON A. J. “An application of recurrent nets to phone probability estimation”. **IEEE Transactions on Neural Networks**, 5(2)1994:298-305.
RYEU J. y CHUNG H. “Chaotic recurrent neural networks and their application to speech recognition”. **Neurocomputing**, 13(2)1996:281-294.
SMYTH P., HECKERMAN D., y JORDAN M. I. “Probabilistic independence networks for hidden Markov probability models”. **Neural Computation**, 9(2)1997:227-269.
TSOI A. y BACK A. “Locally recurrent globally feedforward networks: A critical review of architectures”. **IEEE Transactions on Neural Networks**, 5(2)1994:229-239.
TSOI A. y BACK A. “Discrete time neural network architectures: A unifying review”. **Neurocomputing**, 15(3)1997:183-223.
WAIBEL A. H., HANAZAWA T., HITON G., SHIKANO K., y LANG K. “Phoneme recognition using time-delay neural networks”. **IEEE Transactions on Acoustic Speech and Signal Processing**, 37(3)1989a:328-339.
WAIBEL A. H., SAWAI H., y SHIKANO K. “Modularity and scaling in large phonemic neural networks”. **IEEE Transactions on Acoustic Speech and Signal Processing**, 37(12)1989b:1888-1898.
WANG D., LIU X., y AHALT S. “On temporal generalization of simple recurrent networks”. **Neural Networks**, 9(7)1996:1099-1118.
WATROUS R. “Speaker normalization and adaptation using second-order connectionist networks”. **IEEE Transactions on Neural Networks**, 4(1)1993:21-30.
WIDROW B. y LEHR M. A. “30 years of adaptative neural networks: Perceptron madaline and backpropagation”. **Proceedings of the IEEE**, 78(9)1990:1415-1440.
WILIÑSKI P., SOLAIMAN B., HILLION A., y CZARNECKI W. “Toward the border between neural and Markovian paradigms”. **IEEE Transactions on Systems Man and Cybernetics - Part B: Cybernetics**, 28(2)1998:146-159.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.

WOODLAND P. C. y SMYTH S. G. “An experimental comparison of connectionist and conventional classification systems on natural data”. **Speech Communication**, 9(1)1990:73-82.
WU P., WARWICK K., y KOSKA M. “Neural network feature maps for chinese phonemes”. **Neurocomputing**, 4(1)1992:109-112.
YAMAUCHI K., FUKUDA M., y FUKUSHIMA K. “Speed invariant speech recognition using variable velocity delay lines”. **Neural Networks**, 8(2)1995:167-177.
YAN Y. “Understanding speech recognition using correlation-generated neural network targets”. **IEEE Transactions on Speech and Audio Processing**, 7(3)1999:350-352.
YE H., WANG S., y ROBERT F. “A PCMN neural network for isolated word recognition”. **Speech Communication**, 9(1)1990:141-153.
ZAHORIAN S. A. y NOSSAIR Z. B. “A partitioned neural network approach for vowel classification using smoothed time/frequency features”. **IEEE Transactions on Speech and Audio Processing**, 7(4)1999:414-425.
ZUE V., SNEFF S., y GLASS J. “Speech database development: TIMIT and beyond”. **Speech Communication**, 9(4)1990:351-356.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Reconocimiento automático del habla con redes neuronales artificiales"
Revista Ciencia, Docencia y Tecnología, Vol. XVI, No. 31, pp. 261-322, 2005.