Sistema de Reconocimiento Automático del Habla

Hugo L. Rufiner, Diego H. Milone

Resumen

Este trabajo es el resultado de un emprendimiento multidisciplinario en el que se utilizaron herramientas de campos tales como fonética, fisiología del oído y de la percepción auditiva, lingüística, gramática, redes neuronales, modelización, electrónica analógica y digital, computación e inteligencia artificial. Los objetivos principales son la investigación y el desarrollo de sistemas de reconocimiento automático del habla en idioma español, los cuales puedan brindar una salida de texto escrito para ser utilizado en un procesador de textos o como entrada de comandos en programas convencionales o interfaces gráficas. Un sistema completo para el reconocimiento automático del habla puede ser dividido en una etapa de procesamiento de la señal, una de clasificación de fonemas y una de modelado del lenguaje. El primer bloque realiza un procesamiento de la señal de habla, acondicionándola y realizando un análisis espectral-temporal. Luego, la etapa de clasificación procesa la secuencia de espectros para devolver una cadena de símbolos que corresponden a la secuencia de fonemas o dífonos emitidos. El último módulo consiste en un analizador cuyo objetivo es descifrar las palabras y frases que se hallan contenidas en las cadenas de símbolos pseudo fonéticos que son emitidas por el módulo anterior. En este artículo se revisan las diversas líneas investigadas en torno a estos tres módulos básicos.

Palabras clave

Reconocimiento automático del habla, interfaces hombre-máquina, inteligencia artificial, señal de voz (procesamiento y análisis).

1. Introducción

Uno de los aspectos que mayor importancia ha cobrado en el desarrollo de equipos y programas de computación es la comunicación con el usuario. Hasta el presente se ha hecho uso exclusivo del lenguaje escrito: utilización del teclado y mensajes impresos. Aparejadas a la facilidad de su utilización, las interfases orales ofrecen muchas otras ventajas para la comunicación con la computadora. Una es la velocidad: la mayor parte de la gente puede pronunciar más de 200 palabras por minuto, mientras que muy pocas pueden tipear más de 60 palabras en el mismo tiempo. Mediante el habla también se podrían eliminar algunas de las limitaciones físicas de la actual interacción hombre-máquina: se podría controlar una computadora mientras se trabaja en la oscuridad o sin sentarse al teclado. El reconocimiento automático del habla (ASR: del inglés Automatic Speech Recognition) puede permitirle a los usuarios utilizar una computadora en lugares o situaciones en los que sería imposible o peligroso hacerlo en el modo convencional.

A continuación, como parte de esta introducción se realizará una breve reseña histórica y se revisarán las etapas básicas con que consta un sistema considerado el "estado del arte" en ASR. Luego se introducirá a la complejidad del problema y en base a esto se definirán los distintos tipos de sistemas de ASR que existen en la actualidad. En la Sección 2 de presentarán las diferentes alternativas investigadas, con una breve referencia a los resultados obtenidos en cada caso. En la Sección 3 se describirán brevemente las bases de datos que se utilizaron para evaluar el desempeño de los sistemas de ASR, cerrando en la Sección 4 con las conclusiones globales de este trabajo.

1.1 Breve reseña histórica

Los primeros intentos de desarrollo de sistemas de ASR datan de los años 50. Estos primeros trabajos abordaban el reconocimiento de un vocabulario reducido, del orden de 10 palabras, emitidas por un único locutor. La década de los 60 marca el inicio de tres proyectos que han tenido gran repercusión en el área. Estos proyectos fueron desarrollados por Martin (RCA Labs.), en el campo de la normalización de la voz; Vintsyuck (URSS), en métodos de programación dinámica y Reddy (CMU), que realizó los primeros trabajos en reconocimiento de voz continua. En la década de los 70 se hicieron viables los sistemas de reconocimiento de palabras aisladas. En ésta década, los sistemas de reconocimiento estaban basados en los métodos de programación dinámica. Sin embargo, en los 80 se produce un desplazamiento de estos métodos en favor de los modelos ocultos de Markov (HMM: del inglés Hidden Markov Models), ampliamente utilizados en la actualidad. También se comienzan a utilizar algunas aproximaciones basadas en redes neuronales. En la actualidad existen varios sistemas de ASR, algunos de los cuales están ya siendo comercializados. Se pueden destacar los denominados SPHINX (CMU), BYBLOS (BBN), Dragon Dictate (Dragon Systems) y Naturally Speaking (IBM). La mayoría de estos sistemas están basados en HMM o su versión híbrida con redes neuronales y llegan a un reconocimiento del orden del 95%, en discurso continuo, para un único hablante, con un micrófono de buena calidad y en un ambiente de bajo ruido. A diferencia del ser humano, en otras condiciones el desempeño los sistemas de ASR se degrada rápidamente. Desde el punto de vista práctico esto implica que, a pesar de los avances logrados, aún quedan muchos problemas por resolver para la aplicación masiva de los sistemas de ASR.

1.2 Etapas de un sistema de ASR

La estructura general de los sistemas de ASR tiene esencialmente tres módulos o etapas (Figura 1):

1. Procesamiento o análisis del habla (en inglés se conoce como front-end): en esta etapa se realiza algún tipo de análisis de la señal de voz en términos de la evolución temporal de parámetros espectrales (previa conversión analógica/digital de la señal). Esto tiene por función hacer más evidentes las características necesarias para la etapa siguiente y a veces también limpiar y reducir la dimensión de los patrones para facilitar su clasificación.
2. Clasificación de unidades fonéticas o modelo acústico: esta etapa clasifica o identifica los segmentos de voz ya procesados con símbolos fonéticos (fonemas, difonos o sílabas). A veces se puede asociar una probabilidad con este símbolo fonético, lo que permite ampliar la información presentada al siguiente módulo.
3. Análisis en función de reglas del lenguaje o modelo del lenguaje: en esta última etapa se pueden aprovechar las reglas utilizadas en la codificación del mensaje contenido en la señal para mejorar el desempeño del sistema y producir una transcripción adecuada. Aquí se utilizan otras fuentes de conocimiento como la ortográfica, la sintáctica, la prosódica, la semántica o la pragmática.

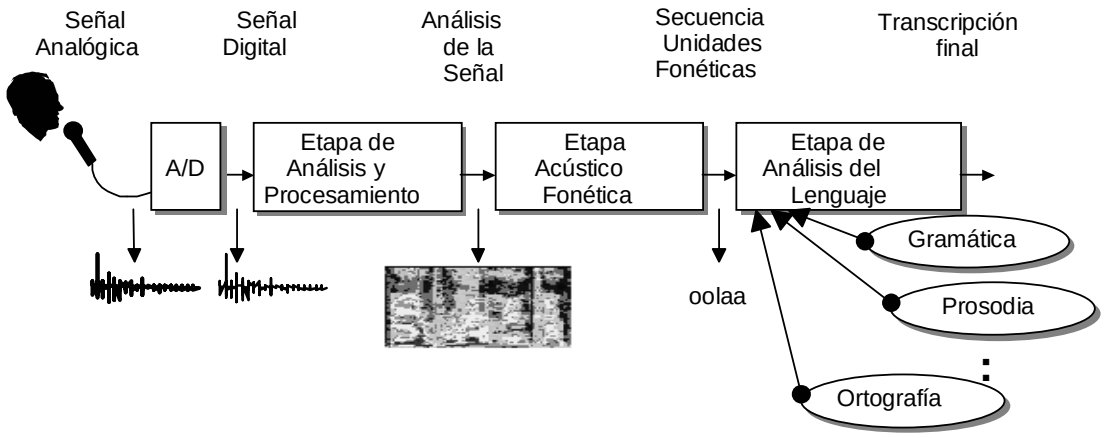

Figura 1: Componentes de un sistema de ASR típico.

1.3 Complicaciones propias del ASR

A pesar de la sencillez que parece presentar el problema del habla para los humanos, el estudio de la misma muestra, de forma inmediata, una enorme complejidad. En ella aparecen mezclados varios niveles de descripción, que interactúan entre si. De esta forma, el problema del ASR presenta una naturaleza interdisciplinaria, y para solucionarlo es necesario aplicar técnicas y conocimientos procedentes de las siguientes áreas (Rabiner y Juang, 1993): procesamiento de señales, física (acústica), reconocimiento de patrones, teoría de la información y comunicaciones, lingüística, fisiología, informática y psicología. Además de la interdisciplinariedad expuesta, existen algunos aspectos prácticos relacionados con el habla que hacen del ASR una tarea difícil. Estos se pueden agrupar en seis categorías (Varile y Zampolli, 1995):
1. Continuidad: en el lenguaje natural no existen separadores entre las unidades, ya que no existen silencios, en algunos casos, ni entre las palabras.
2. Dependencia del contexto: cada sonido elemental en los que se puede dividir el habla (fonema) es modificado por el contexto en el que se encuentra. De esta forma, se produce el efecto denominado coarticulación, según el cual los fonemas anterior y posterior a uno dado modifican el aspecto del mismo. Aparecen también efectos de orden superior, dependiendo la pronunciación de un fonema, de su situación en una palabra o incluso en una frase.

3. Variabilidad: se pueden distinguir dos tipos de variabilidad. La variabilidad intra-hablante está relacionada con las modificaciones introducidas por un mismo hablante sobre diferentes pronunciaciones de los mismos fonemas o palabras. Incluso en idénticas condiciones, cada pronunciación presentará diferencias con las restantes debido a la diferente duración temporal. La variabilidad inter-hablante se debe a aspectos relacionados con el locutor y el entorno, ya que la señal obtenida dependerá de los dispositivos utilizados en su captación, del entorno donde se obtiene y, principalmente, de aspectos anatómicos particulares del aparato fonador de cada hablante.
4. Necesidades de almacenamiento: debido a las causas anteriores, se hace necesario procesar y almacenar grandes cantidades de datos.
5. Estructuración: la misma señal contiene información sobre varios niveles de descripción. De esta forma, una frase puede ser descrita a nivel semántico, sintáctico o fonético. Por otra parte, una señal de voz contiene información sobre el locutor que la emite. Así, es posible distinguir el sexo y la identidad de la persona a partir de la propia señal. Un sistema de ASR tendría que determinar qué información de las citadas es de interés para lograr su objetivo.
6. Inexistencia de reglas de descripción y redundancia: no existen reglas precisas capaces de describir los diferentes niveles en los que se presenta la información. Es más, cada uno de los niveles citados anteriormente aparece fuertemente relacionado con los demás, dificultando el análisis de la voz.

1.4 Tipos de reconocimiento

Los factores expuestos en el apartado anterior hacen inviable todavía el poder abordar el problema del ASR de forma global. Se hace necesario, por consiguiente, establecer hipótesis simplificadoras que restrinjan el campo de aplicación del sistema. De esta forma se introducen restricciones sobre diferentes aspectos de la señal de voz a reconocer, como pueden ser el número y tipo de los locutores, tamaño del vocabulario, etc. Atendiendo a las restricciones impuestas, los sistemas de ASR pueden clasificarse según varios criterios.
1. En función del hablante: según los grupos de locutores que se utilizan en el entrenamiento y evaluación del sistema, se pueden distinguir tres tipos de reconocedores. En los reconocedores monolocutor el sistema se entrena y evalúa para un único hablante. En los reconocedores multilocutor el entrenamiento y evaluación se realiza sobre el mismo conjunto de locutores. Finalmente, en los reconocedores independientes del hablante la evaluación del sistema se efectúa sobre un conjunto de locutores diferente del utilizado para entrenarlo.
2. En función de la manera de hablar y la aplicación: de acuerdo con el tipo de locución requerido para el funcionamiento adecuado del sistema existen los siguientes seis tipos. En los reconocedores de palabras aisladas (IWR, Isolated Word Recognition) el reconocimiento se realiza sobre palabras completas emitidas de forma aislada entre sí. En los reconocedores de palabras conectadas (CWR, Connected Word Recognition) se utiliza también a las palabras como unidades de reconocimiento, pero éstas pueden ser emitidas secuencialmente con pausas entre ellas. Los reconocedores de voz o discurso continuo (CSR, Continuous Speech Recognition) realizan la tarea atendiendo a unidades inferiores a la palabra (fonemas, difonos, etc.) y sobre frases completas, sin necesidad de establecer silencios entre las palabras que las constituyen. Los reconocedores de voz para programas de noticias deben trabajar con habla espontánea pero producida por un locutor profesional en algún tipo de estudio de radio o televisión. En los reconocedores de palabras clave (Word Spotting) la locución suele ser cuidadosa, aunque no es una condición ya que su finalidad es la detección (o reconocimiento) de las palabras clave contenidas en el vocabulario. Los reconocedores de habla espontánea (SSR, Spontaneous Speech Recognition) buscan resolver el caso más complicado, ya que aquí

suelen no respetarse algunas reglas del lenguaje, e inclusive es muy probable que en la señal aparezcan eventos acústicos diferentes al habla, como la tos, el estornudo, el hipo, etc.
3. En función de las distorsiones de la señal: también se pueden tener en cuenta las distorsiones producidas en la señal de voz por distintos factores. De esta forma podemos separar a los reconocedores en dos tipos. Los reconocedores de habla limpia son aquellos entrenados y probados en condiciones de laboratorio mientras que los reconocedores robustos (RSR, Robust Speech Recognition) permiten su utilización en ambientes reales, donde varios factores complican la tarea (ruido, reverberación, canal de transmisión, etc.).

2. Avances en reconocimiento automático del habla

Para este trabajo la metodología general consistió en la propuesta, ajuste y evaluación sistemática de diferentes alternativas para cada una de las etapas del sistema, según se describió anteriormente. Se realizaron experimentaciones numéricas en computadora para determinar el desempeño de cada una de estas alternativas. Esto se completó con un registro exhaustivo y sistemático de los resultados de dicha experimentación, para poder compararlos con los obtenidos en cualquier etapa posterior y con el "estado del arte" en ASR. Para el desarrollo de este artículo se eligieron las técnicas que mostraron un mejor desempeño para cada etapa, teniendo también en cuenta la fiabilidad, el costo de implementación y la velocidad. A continuación describimos sintéticamente cada una de ellas.

2.1 Etapa de procesamiento del habla

En la práctica, la mayoría de las señales se encuentran en el dominio del tiempo. Esta representación no siempre es la más apropiada cuando se tiene por objetivo su procesamiento para el ASR. En muchos casos, la mayoría de la información distintiva se encuentra oculta en el contenido frecuencial de la señal o en alguna otra forma de representación. La etapa de procesamiento convierte la señal de voz en algún tipo de representación paramétrica (generalmente con una cantidad de información considerablemente menor pero significativa) para su análisis posterior. Se supone que cuanto mejor sea el proceso utilizado para generar los patrones a utilizar, estos son más fáciles de identificar por el clasificador. Probablemente, la representación paramétrica más importante de la voz es la envolvente espectral de corta duración. Por lo tanto, los métodos de análisis espectral son generalmente considerados el núcleo de la etapa de procesamiento de señales en un sistema de ASR. Más recientemente comenzaron a considerarse enfoques basados en otro tipo de representaciones y son a los que se ha dado mayor énfasis en este trabajo. Las técnicas de procesamiento del habla exploradas pueden resumirse en:

1. Análisis Tradicional:
   1.1 Fourier (STFT, Short Time Fourier Transform)
   1.2 Cepstra (CC, Cepstral Coefficients)
   1.3 Mel cepstra (MFCC, mel Frequency CC)
   1.4 Delta mel cepstra (MFCC+D)
2. Modelos de oído
3. Análisis mediante onditas (WT, Wavelet Transform):
   3.1 Transformada ondita Discreta Diádica (DWT, Discrete WT)
   3.2 Paquetes de onditas (WP, Wavelet Packets)
   3.3 Paquetes de onditas orientados perceptualmente (POWP, Perceptualy Oriented WP)
   3.4 Paquetes evolutivos de onditas (EWP, Evolutive Wavelet Packets)
4. Representaciones ralas (SR, Sparse Representations): WP como diccionarios para lograr SR (WPSR)
5 Análisis no lineal: medidas de complejidad y entropía (CM, Complexity Measures).

6. Técnicas de Adaptación:
  6.1 Filtrado óptimo probabilístico (POF, Probabilistic Optimum Filtering).
  6.2 Filtrado mediante redes neuronales (ANNF, Artificial Neural Networks Filtering).

2.1.1 Procesamiento tradicional

Las técnicas de procesamiento tradicional fueron estudiadas e implementadas a lo largo de todas las etapas de la investigación. A partir de la mejor de todas las alternativas evaluadas (MFCC+D) se construyó la etapa de procesamiento para el sistema de referencia utilizado en los experimentos.

2.1.2 Modelos de oído

El objetivo de esta exploración consistió en encontrar una técnica de procesamiento óptima que permita conseguir un mecanismo de codificación (de la información contenida en la señal de voz) robusto e independiente de características individuales de los hablantes. En estudios preliminares se propusieron las primeras alternativas para las diferentes etapas de un sistema de ASR: transformada de Fourier para la etapa de procesamiento, redes neuronales con retardos para la acústico-fonética y sistemas expertos para la de análisis del lenguaje. Sin embargo, estas etapas no se integraron completamente y el enfoque tenía problemas para ser escalado a situaciones más complejas que la planteada originalmente. Posteriormente se encaró un enfoque más biológico de la etapa de procesamiento, utilizando modelos de oído (Rufiner, 1994). Aunque el tiempo de cálculo de estos últimos resultaba prohibitivo, su semejanza con el análisis basado en onditas y las atractivas características teóricas de estas últimas nos llevaron a explorar en años posteriores este tipo de análisis.

2.1.3 Onditas

Las onditas aparecen a principios de los 90 constituyendo una alternativa atractiva por su carácter intrínsecamente transitorio. Sin embargo los intentos por superar definitivamente a las técnicas del estado del arte no han sido provechosos todavía, por lo menos en lo que hace a su integración con sistemas basados en HMM. Aunque posteriormente volveremos sobre este tema, se debe recalcar que los módulos no son tan independientes como se puede pensar en principio, por lo tanto los resultados dependen en realidad de todas las etapas implicadas, su interacción y su ajuste. El escaso éxito de las onditas en este campo queda demostrado parcialmente por la escasa cantidad de publicaciones en que emplea esta herramienta para el ASR. Entre 1991 y 1994 aparecieron sólo algunos pocos trabajos, la mayoría de ellos en anales de congresos. En muchos de estos trabajos se reportan algunas mejoras con respecto al procesamiento tradicional, que no suelen ser significativas, y por lo tanto esta técnica no logra desplazar a los enfoque más clásicos (como MFCC). Los resultados de nuestras investigaciones coinciden con los obtenidos por la comunidad de ASR.
En (Rufiner, 1996) se realizó la comparación entre el análisis basado en onditas y el basado en la transformada de Fourier. A pesar de sus atractivas cualidades para el procesamiento de señales transitorias, la comparación fue desfavorable para las onditas (en realidad se exploró principalmente el caso de la DWT, en contraste con la STFT). Durante esta comparación se advirtió la multiplicidad de facetas que presentaba este tipo de representaciones relacionadas con onditas (elección del tipo de transformación, familia de onditas, parámetros, etc.), y la dificultad de brindar una respuesta definitiva acerca de cuál es la mejor alternativa para esta aplicación. Posteriormente se continuó la exploración de la aplicación de onditas, tratando de superar algunos de los problemas detectados, como el de conseguir una adecuada resolución frecuencial en la zona de las frecuencias bajas. Para ello se recurrió a diferentes aproximaciones derivadas de los WP que mejoraron notablemente el desempeño de esta etapa con respecto a la DWT (Gamero y Rufiner, 1998; Torres y Rufiner, 1998). Los WP, son

considerados como una extensión del clásico análisis multiresolución (MRA). Este esquema MRA, propuesto por Mallat (1999), ha sido ampliamente utilizado en aplicaciones tales como: compresión, eliminación de ruido, caracterización de singularidades y detección de no estacionariedades en el análisis de señales e imágenes. Entre estas aproximaciones desarrolladas por nuestro grupo alrededor de WP podemos citar:

- Paquetes de onditas orientadas perceptualmente: se diseño un banco de filtros de manera que la base realice un análisis similar al que realiza el oído.
- Paquetes evolutivas de onditas: esta estrategia constituye una alternativa que permite encontrar la base óptima para un problema de clasificación determinado.

Luego del ajuste de distintos parámetros las pruebas con algunas familias de onditas como Symmlets, y en particular, para los experimentos realizados con POWP, se obtuvieron resultados comparables a los obtenidos mediante MFCC+D (aunque no se ensayaron todos los casos posibles). Los EWP son muy promisorios pero requieren todavía la solución de algunas cuestiones de implementación práctica ya que necesitan de mucho tiempo de cálculo para lograr resultados. En este sentido se modificó el algoritmo genético desarrollado originalmente de forma tal de maximizar las distancias entre fonemas (en lugar de utilizar directamente un clasificador que requería más recursos de cómputo). Los parámetros a optimizar fueron: la utilización o no de la integración por bandas, el coeficiente de pre-énfasis, el tipo de ventana, la familia de onditas, parámetros de la ondita, y el árbol de filtros de la WP. De esta forma se pudieron atacar problemas un poco más grandes pero todavía no del tamaño necesario para nuestra aplicación. Posteriormente se continuó con el trabajo simplificando algunas hipótesis, entre ellas la necesidad de una representación ortogonal (Sigura, 2001). La restricción acerca de la ortogonalidad es en realidad un tanto forzada por consideración de simplificación matemática y se había venido respetando históricamente, a pesar de no existir evidencias físicas o fisiológicas de su necesidad.

2.1.4 Representaciones ralas

Se puede decir que el enfoque tomado en nuestros trabajos sobre la etapa de procesamiento está sin duda influenciado por la teoría y las ideas detrás de la WT. Una de las actividades actuales es extender los resultados anteriores mediante WP e investigar sobre la posibilidad de encontrar una base, no necesariamente ortogonal, que maximice algún criterio (clasificación, dispersión, independencia, etc). Recientemente se han encontrado conexiones interesantes entre el análisis de señales mediante bases sobrecompletas y la manera en la que el cerebro parece procesar algunas señales sensoriales. Este tipo de análisis puede dar representaciones que poseen muy pocos elementos activos o diferentes de cero, o sea sumamente ralas. Esta es una característica que comparten los sistemas sensoriales biológicos, y hace que la información sea fácilmente codificables en términos de trenes de pulsos o espigas. Existen varios métodos que permiten encontrar una representación rala de una señal si se posee una base sobrecompleta adecuada. Así mismo se pueden diseñar los elementos de estas bases a la medida del tipo de "pistas" que se pretenden encontrar en la señal, o buscar automáticamente la base o diccionario que satisfaga algún criterio particular. Las representaciones ralas poseen también una robustez intrínseca al ruido aditivo, cuando este no pueda ser expresado fácilmente en términos de los elementos de la base. Recientemente hemos iniciado el estudio del comportamiento de estas técnicas para señales sintéticas (Rufiner y cols., 2001; Rufiner y cols., 2002), pero hay relativamente poco trabajo realizado en señales del mundo real.

2.1.5 Medidas de complejidad

Otra línea explorada fue la aplicación de técnicas de análisis no lineal de señales al caso del habla. La idea original surge a partir de publicaciones que sugieren que el mecanismo de producción de la voz podría ser un proceso no lineal. Este hecho ha generado gran interés en el

área y ha inducido la investigación de la existencia de atractores caóticos de baja dimensión en la voz. En particular se trabajó sobre las relaciones entre diagrama espectral y caos, y en la reconstrucción de un atractor dada una serie temporal. Varios trabajos de índole general que se desarrollaron en conjunto con integrantes del Laboratorio de Dinámicas no lineales y Señales (Torres y cols., 1995; Gamero y cols., 1997), fueron dando las bases para la aplicación directa de estas ideas en el campo del ASR. Finalmente se implementaron e integraron diferentes medidas de complejidad a un sistema de ASR del "estado del arte" basado en MFCC y HMM, demostrándose así el aporte de esta información para mejorar el ASR en condiciones de ruido elevado (Rufiner y cols., 2004). También se utilizaron CM para segmentar el habla según se explica más adelante.

2.1.6 Técnicas de adaptación

El filtrado óptimo probabilístico constituye una técnica de limpieza de ruido en el dominio de las características extraídas de la señal de habla continua, con el fin de implementar sistemas de RSR para habla continua. El POF constituye un mapeo multidimensional lineal por tramos entre el espacio de las características de las señales ruidosas y el de las señales limpias. Para la estimación del filtro probabilístico se que requiere tener muestras apareadas de las señales con y sin ruido. Las ANN constituyen aproximadores universales de funciones arbitrarias y por lo tanto se convierten en una alternativa interesante para la solución de este problema. Estas estructuras permiten el mapeo directo entre ambos espacios de manera no lineal, lo que constituiría un verdadero filtro no-lineal. Se realizó una comparación entre ambos métodos sobre un conjunto de datos tomados del Corpus de habla continua en español LATINO 40 (ver Sección 3). Estos datos se mezclaron con ruido blanco obtenido de la base NOISEX en diferentes proporciones. Las señales filtradas mediante ambos métodos se pasaron a través de un reconocedor basado en HMM continuos de mezclas gaussianas, entrenados con habla limpia de la misma base de datos, para confrontar los niveles de reconocimiento de ambas técnicas y compararlos también con los casos extremos (habla limpia y habla ruidosa). En estos experimentos POF se comportó mejor que las ANN (Rufiner y cols., 2000), aunque existen trabajos recientes donde luego de cambiar la regla de aprendizaje de estas últimas se logran muy buenos resultados.

2.2 Etapa de clasificación de unidades fonéticas

Como se explicó anteriormente, el objetivo de esta etapa consiste en obtener una representación de la señal de voz como una cadena de símbolos asociados con los eventos acústico-fonéticos. Para ello existen varias alternativas, de las cuales en la presente investigación se analizaron las siguientes:

1. Técnicas tradicionales:
   1.1 Modelos ocultos de Markov (HMM, Hidden Markov Models)
   1.2 Árboles de decisión (DT, Decision Trees)
2. Redes neuronales artificiales:
   2.1 Perceptrones multicapa (MLP, Multi-Layer Perceptrons)
   2.2 Redes neuronales con retardos temporales (TDNN, Time Delay NN)
   2.3 Redes neuroanles recurrentes (RNN, Recurrent NN)
3. Otras:
   3.1 Máquinas con soporte vectorial (SVM, Support Vector Machines)
   3.2 Técnicas híbridas

Dado que gran parte de los trabajos iniciales se realizaron en torno la utilización de ANN para el ASR, se puede consultar una extensa revisión del tema en Milone, 2001. A continuación

presentaremos la forma en la que se exploró cada técnica tratando de seguir un orden lógico en el tratamiento.

2.2.1 Obtención de un MPL a partir de un DT

Dos de los métodos de aprendizaje maquinal más ampliamente utilizados en tareas de clasificación son los DT y las ANN. En este último caso la arquitectura más utilizada son los perceptrones multicapa. A partir de experiencias previas en la utilización de MLP para reconocimiento de fonemas (Rufiner, 1994), en esta primer etapa se analizaron los distintos métodos de creación de DT y un método para implementar un MLP a partir de un DT. Se comparó el desempeño de este MLP en relación al DT original y con respecto a un MLP definido por separado. Los resultados permitieron definir de manera eficiente la arquitectura de un MLP que actúa como clasificador. Este hecho es de suma importancia ya que no se ha resuelto de manera teórica cómo definir de óptimamente la arquitectura, y la alternativa de prueba y error consume muchos recursos computacionales, especialmente para problemas complejos como el nuestro (Goddard y cols., 1995). Se determinó la importancia del análisis realizado por los DT en la determinación de la estructura del MLP, concluyendo que dan una buena aproximación a la arquitectura óptima, así como también las limitaciones del podado de los DT en cuanto a la generalización.

2.2.2 Redes neuronales con retardos temporales

En la sección de procesamiento se describió la utilización de la transformada WP como análisis de la señal de voz para un sistema de ASR. Cuando se validó esta propuesta el clasificador consistía en una TDNN. En estos trabajos se exploró cómo se comportaban los distintos POWP, tomando el logaritmo de la energía de cada banda como patrón de entrada para el clasificador, el cual consistía en una sola gran TDNN. Adicionalmente se propuso y ensayó una estructura de TDNN modular, la cual consistió en una red para cada una de las clases (fonemas) presentes. Esto proveyó un significante aumento en los porcentajes de reconocimiento del sistema. Los experimentos se realizaron con las oraciones correspondientes a la región 1 de la base de datos de voz TIMIT (descripta más adelante), sobre los fonemas: /b/, /d/, /eh/, /jh/ y /ih/, los cuales fueron también utilizados otros experimentos. Los patrones para el entrenamiento se obtuvieron utilizando la transformada WP, con las onditas Splines y Daubechies, que son las que mejores resultados habían producido hasta ese momento. El sistema así conformado obtiene aumentos relativos en reconocimiento del orden del 10% respecto a un sistema con un clasificador constituido por una única TDNN (Torres y Rufiner, 1998; Rufiner y cols., 2000).

2.2.3 Tecnicas híbridas de clasificación

En esta línea se exploró la posibilidad de utilizar redes neuronales modulares jerarquizadas para el reconocimiento de fonemas. En este contexto, se utilizó un enfoque híbrido con DT. La alternativa propuesta a los DT clásicos consiste en la utilización de ANN para la implementación de la tarea de decisión realizada en cada nodo. El aporte realizado en este trabajo consistió en el diseño de una estrategia novedosa para el crecimiento "inteligente" de los DT basados ANN (Milone y cols., 1998a). Se realizó, además, un programa de computadora que implementó tal estructura de clasificación mediante un árbol de redes neuronales no supervisadas. El software permite la visualización de la estructura del árbol generado, así como los parámetros calculados que controlan el crecimiento de este. También se implementaron diversas herramientas que permiten monitorear el proceso de entrenamiento del árbol. Los entrenamientos se realizaron con una base de datos que contiene la posición frecuencial de las tres primeras formantes de 660 vocales inglesas (Milone y cols., 1998b).

2.2.4 Modelos probabilísticos basados en mezclas gaussianas

Iniciando la aplicación de los modelos probabilistas a esta etapa se implemento un modelo de mezclas gaussianas para la distribución de las vocales del español en el espacio formántico. Para ello se calcularon los valores de las formantes de las vocales aisladas del español rioplatense pronunciadas por adultos normoyentes. El objetivo de este estudio fue también el de obtener patrones formánticos para ser utilizados como normativa en estudios de voces normales y patológicas. Se realizaron registros de 40 voces femeninas y 40 voces masculinas correspondientes a sujetos entre 18 y 35 años, nativos de habla español rioplatense sin ningún tipo de patología vocal asociada. El análisis de los 3 primeros formantes y los respectivos anchos de banda, así como el 4to formante, cuando fue posible, se realizó utilizando estimadores espectrales. Los contornos formánticos se midieron mediante el método de predicción lineal (LPC: Linear Predictictive Coding). Se encontraron resultados similares con respecto a estudios previos de otros autores, a pesar de diferir en los métodos de análisis, cantidad de sujetos participantes y los casi 25 años que separan los estudios. Por otra parte se pudo concluir que el modelo era adecuado para este tipo de datos (Aronson y cols. 2000).

2.2.5 Utilización de HMM continuos

Se ha avanzado mucho en el ASR mediante la aplicación de los HMM principalmente porque esta técnica ha permitido modelar la gran variabilidad del habla. Este enfoque modela la señal de habla como un proceso doblemente estocástico que ocurre a través de una red de estados interconectados. Cada uno de estos estados representa algún aspecto importante de la señal de habla. Un HMM es una "máquina de estados finitos" capaz de generar secuencias de observaciones que son asociadas con alguna unidad del habla (Deller y cols., 1993). El HMM es entrenado a fin de que pueda representar fielmente el modelo estadístico de las secuencias de observación para una unidad determinada (palabra, sílaba o fonema). Para la realización de los experimentos se utilizó el toolkit denominado HTK, que implementa los algoritmos mencionados y una serie de utilidades para el diseño de sistemas de ASR basados en HMM. Se entrenó un HMM continuo utilizando el procesamiento tipo MFCC+D, incluyendo también la energía. Los resultados se compararon con resultados obtenidos con las estrategias de redes neuronales y diferentes técnicas de preprocesamiento (Fourier, Fourier en escala de mel, diferentes familias de Wavelets, etc). Estos últimos superaron ampliamente los obtenidos con las otras técnicas de clasificación. Sin embargo se debe destacar que en el caso de HMM se utiliza un modelo por cada fonema, por lo que para que la comparación sea "justa" debería utilizarse una red neuronal para cada uno de los fonemas por separado. Esto sugirió la realización de los experimentos con TDNN modulares ya comentados, donde se obtuvieron mejoras importantes,

2.2.6 HMM vs RNN

Las TDNN pueden aproximar dinámicas finitas y se exploraron en los primeros trabajos. Por otra parte las RNN tienen capacidades teóricas de aproximación de sistemas dinámicos en general. Por lo anterior se investigaron las diferentes arquitecturas de RNN existentes a fin de realizar una taxonomía que permitiera elegir la más adecuada para el ASR. Esta taxonomía sirvió de base para establecer relaciones entre las redes recurrentes y los modelos ocultos de Markov (véase Milone, 2001). La idea de esta comparación consistía en encontrar relaciones entre las RNN y los HMM que proporcionaran una visión más global y unificada de los modelos utilizados para resolver el problema del ASR. Esto permitiría encontrar mejores soluciones y aumentar el desempeño de los sistemas existentes en problemas difíciles de resolver con las técnicas actuales, como la robustez al ruido o la independencia del hablante. Como resultado de este estudio se concluyó que, si bien la capacidad teórica de las RNN les permite aproximar el comportamiento de sistemas dinámicos arbitrarios y existen numerosas conexiones con la

teoría de autómatas deterministas y probabilistas, no existen en la actualidad técnicas adecuadas para su entrenamiento en una gran variedad de situaciones. En particular en nuestro caso existe un problema de atenuación exponencial de la memoria de la red que le impide aprender relaciones entre eventos separados en el tiempo. Más recientemente han despertado cierto interés de la comunidad de ANN la redes pulsadas con sinápsis dinámicas (PNN). La aplicación de PNN al ASR podría constituir una linea de investigación futura.

2.3 Análisis en función de reglas del lenguaje

Esta es la etapa que menos se investigó en nuestro trabajo, sin embargo entre las técnicas exploradas podemos mencionar:

1. Métodos tradicionales: bigramáticas suavizadas por "backing-off".
2. Incorporación información prosódica a un sistema de ASR basado en HMM.
3. Sistemas de diálogo

2.3.1 Prosodia, acentuación y ASR basado en HMM

En los sistemas de reconocimiento automático del habla se han ido incorporado gradualmente los diferentes aspectos relacionados tanto con la producción como con la percepción natural del habla. El conjunto de características prosódicas del habla no ha sido utilizado hasta el momento de una forma explícita en el proceso mismo de reconocimiento. Hemos realizado un análisis de los tres parámetros más importantes de la prosodia: energía, entonación y duración (Milone y Rubio, 2001) y propuesto un método para incorporar esta información al ASR (Milone y cols. 2001). A partir del análisis preliminar se diseño un clasificador de rasgos prosódicos que asocia estos parámetros con la acentuación ortográfica. La información de acentuación es incorporada por medio del modelo de lenguaje en un reconocedor estándar basado en HMM. Se realizaron varios experimentos donde se impusieron diversos grados de dificultad a la hora de realizar las pruebas de validación. Los resultados finales proveyeron una reducción porcentual del error que llegó hasta el 43.00% (Milone y Rubio, 2003).

2.3.2 Sistemas de diálogo

En trabajos recientes se puede encontrar un considerable número de sistemas automáticos destinados a proporcionar diversos tipos de servicios a los usuarios haciendo uso del habla. Estos sistemas utilizan principalmente las tecnologías de reconocimiento y comprensión del habla, control del diálogo, y generación de voz. Sin embargo, los sistemas automáticos que interactúan a partir de la voz deben tratar diversos problemas que no son resueltos por el reconocedor, como las diferencias en la voz de los usuarios, falso comienzo de las frases, ruido ambiental, cruce de conversaciones, palabras no incluidas en el vocabulario, etc. En colaboración con el Grupo de Investigación en Procesamiento de Señales y Comunicaciones de la Universidad de Granada se estudiaron algunos problemas relacionados con el funcionamiento en tiempo real de los sistemas automáticos de diálogo (López-Cózar y cols., 2000). El trabajo se centra en una metodología utilizada para intentar que el tiempo de respuesta de un sistema de diálogo sea aceptable por los usuarios. A partir de los resultados obtenidos se puede concluir que la estrategia utilizada es aceptable para la mayoría de las tareas de reconocimiento consideradas. Asimismo, los resultados muestran que es necesario realizar cambios en la estrategia usada para otras tareas ya que la tasa de comprensión es insuficiente y el tiempo de reconocimiento es excesivo para un sistema interactivo (López-Cózar y Milone, 2001).

2.4 Segmentación automática de señales de voz

La segmentación de voz consiste en dividir una emisión en diferentes trozos de acuerdo con algún criterio. Es común que se realice segmentación de voz para separar en fonemas y también suele ser de interés la segmentación según sílabas o unidades de nivel superior, como la palabra. Si bien la segmentación constituye una parte esencial en todo sistema de ASR, es difícil ubicarla entre alguna de las etapas descriptas anteriormente. Los HMM realizan una segmentación durante el mismo proceso de búsqueda de la secuencia más probable y utilizan para esto información de los vectores de voz preprocesados, de los modelos acústicos e incluso de los modelos de lenguaje. Es así como puede verse al proceso de segmentación como un componente fundamental del ASR en si mismo pero no como un etapa con una ubicación más o menos específica. Por otro lado, cuando se utilizan redes neuronales estáticas (como un MLP), es requisito indispensable tener una segmentación previa para poder luego entrenar y probar el sistema. La tarea de segmentación y etiquetado de la señal de voz es bastante ardua, y el hecho de contar con una presegmentación mediante algoritmos automáticos permite disminuir considerablemente el tiempo empleado para completarla. Con este objetivo se implementaron enfoques diferentes para abordar la solución del problema:

1. Método tradicional: HMM con alineación forzada.
2. Análisis de cambios de entropia multiresolución.
3. Redes Neuronales.
4. Algoritmos Evolutivos.

2.4.1 HMM con alineación forzada

Esta es la alternativa automática más difundida. La principal desventaja de este método es que hay que entrenar un sistema de ASR basado en HMM por completo para luego poder hacer una segmentación. En lugar de utilizar este sistema como reconocedor, se alimenta al sistema con la transcripción (texto) y la grabación de voz, y se busca la secuencia de estados más probable (algoritmo de Viterbi). Una vez obtenida la secuencia de estados más probable la segmentación queda completamente definida.

2.4.2 Análisis de cambios de entropía multirresolución continua

En este desarrollo utilizamos la entropía multirresolución continua (CME) la cual consiste básicamente en calcular la entropía en una representación de los datos mediante onditas. Teniendo en cuenta que, ante la presencia de cambios de complejidad en la señal la CME presenta variaciones abruptas en su valor, coincidentes con la localización temporal de dicho cambio, se implementó un detector automático el cual combina la CME con las técnicas de análisis de componentes principales y de detección de cambios abruptos de tipo estadístico. Se obtuvieron resultados en los que las marcas generadas por el detector coincidieron relativamente con transiciones de palabras y fonemas, si bien no exactamente y no para todos estos. Teniendo en cuenta que la CME fue diseñada como una herramienta útil para detectar cambios en complejidad en señales no lineales, no estacionarias, los resultados obtenidos sugieren que ciertas transiciones de palabras y fonemas podrían corresponder a cambios de complejidad en los modelos subyacentes (Torres y cols., 2003a). Por otra parte, aparece como posible que, con una correcta selección de la ondita y de los distintos parámetros libres de la CME y del detector pueda desarrollarse una herramienta que permita realizar una segmentación automática de cierto número de fonemas. Sin embargo, este último aspecto deberá ser motivo de un estudio mas exhaustivo.

2.4.3 Segmentación mediante redes neuronales

Se implementaron redes neuronales tipo MLP para segmentar automáticamente los fonemas en base a dos parámetros temporales sencillos: energía y cruces por cero. Para agregar robustez a la tarea de clasificación, se calcularon también la derivada primera y la derivada segunda de

cada uno de los parámetros. Los MLP se entrenaron con el algoritmo de retropropagación y se probaron distintas arquitecturas, manteniendo siempre 6 neuronas en la capa de entrada y 4 neuronas en la capa de salida. En la capa oculta se probó con 10 y 15 neuronas. Los resultados obtenidos en la evaluación del desempeño de las redes no fueron satisfactorios. Esto puede deberse a que las características utilizadas para confeccionar los patrones no sean suficientes para detectar los cambios de dinámica que se producen en las transiciones fonémicas. Por ello, en los experimentos futuros deberán elegirse otras características. El empleo de la entropía multiresolución según se describió en la sección anterior podría ser una alternativa.

2.4.4 Segmentación mediante un algoritmo evolutivo

Los métodos de computación evolutiva (EC, Evolutionary Computation) manipulan una población de soluciones potenciales codificadas en cadenas o vectores que las representan. Este conjunto de cadenas representa el patrimonio genético de una población de individuos. Los operadores artificiales de selección, cruza y mutación son aplicados para buscar los mejores individuos (mejores soluciones) a través de la simulación del proceso evolutivo natural. Cada solución potencial se asocia con un valor de aptitud, que mide que tan buena es comparada con las otras soluciones de la población. Este valor de aptitud es la simulación del papel que juega el ambiente en la evolución natural darwiniana. En este trabajo se utilizó un algoritmo basado en EC para la segmentación de voz. Las pruebas que se realizaron se dividen en dos partes. En primer lugar están las pruebas que tienden mostrar las características más importantes del algoritmo. Esto se realiza mediante un secuencia que no es de voz sino que ha sido creada artificialmente y contiene información que resulta en una segmentación obvia. Las segundas pruebas se realizaron en archivos de voz reales y se comparan resultados con el proceso de segmentación manual y la segmentación realizada por HMM. Los resultados demostraron una correcta segmentación, bastante similar tanto a la ideal (realizada a mano) como a la realizada usando HMM. El algoritmo evolutivo para la segmentación de voz se presenta como un método útil para segmentar bases de datos donde la cantidad de segmentos por frase se conoce de antemano (Milone y Merelo, 2002). Las diferentes modalidades de segmentación mostradas parecen ajustarse a la segmentación ideal y sería posible optimizar el tiempo de cálculo y corregir algunos errores inducidos con la dependencia de la energía de la señal presente el vector de voz parametrizada. En esta primera realización del algoritmo se dejan abiertas varias posibilidad para realizar diferentes mejoras tanto en lo relativo a la evolución como en el procesamiento de la señal que no se restringe necesariamente a voz. De hecho, tanto el formalismo como la implementación son totalmente aplicables a cualquier otra señal unidimensional.

3 Datos y validación de sistemas de ASR

En esta sección discutimos brevemente los datos utilizados para la experimentación. El énfasis está puesto en las bases de datos desarrolladas por nuestro grupo, dado que el resto son ampliamente conocidas en el medio.

3.1 Bases de datos

1. Peterson: esta es una pequeña base pública con datos sobre las formantes del inglés que se utilizó en algunas pruebas iniciales de los clasificadores.
2. TIMIT: esta base de datos es la más utilizada en el campo del ASR, está formada por unas 6000 frases en idioma inglés. Fue la base utilizada en las primeras pruebas.

3. TIMEX: se realizaron tareas conjuntas con la UAM (México) a fin de lograr el corpus de fonemas en castellano que culminó en el diseño, la grabación y etiquetado de un corpus formado por 10 hablantes mexicanos. Las tareas de etiquetado se llevaron a cabo en el Laboratorio de Audiología de la UAM. Se trabajó sobre estas grabaciones etiquetadas de tal manera de adaptarlas y organizarlas en forma similar al corpus TIMIT (inglés), ya que todas las rutinas de análisis y clasificación se desarrollaron en base a esta estructura. Para esto, primero se submuestrearon las grabaciones, que originalmente se habían adquirido a 22050 Hz, obteniéndose archivos de señales muestreadas a 16000 Hz. Una vez organizada la base de oraciones, se procedió a la obtención de patrones de entrenamiento y prueba para cinco fonemas, a partir de las rutinas para procesamiento con MFCC, de manera de poder realizar una comparación con los resultados obtenidos para los cinco fonemas más cercanos de TIMIT. Estos patrones se utilizaron para entrenar una red neuronal tipo TDNN, obteniéndose resultados preliminares muy buenos (mejores que para TIMIT) aunque poco significativos debido al pequeño numero de hablantes de este corpus.
4. Latino 40: esta base esta formada por 40 hablantes latinos y desarrollada por el Stanford Research Institute (SRI). Fue la primera base en castellano de tamaño mediano utilizada en el PID. Los experimentos con POF se realizaron con esta base. Un problema que presenta es la gran variabilidad dialéctica de los datos.
5. Albayzin: esta es una base en idioma español desarrollada por un grupo de universidades españolas. Se ha podido tener acceso a la misma por colaboración con el Grupo de Investigación en Procesamiento de Señales y Comunicaciones de la Universidad de Granada y se describe adecuadamente en la separata sobre el sistema de referencia, ya que fue utilizada para su desarrollo.
5. CSLU: recientemente se adquirieron una serie de datos pertenecientes a conversaciones telefónica en múltiples idiomas que permiten atacar los problemas de degradación qu e presenta este canal.

3.2 Obtención de resultados

Para la obtención de las medidas finales de desempeño de los sistemas se debe estimar el error cometido en alguna tarea particular. Para ello el método clásico en bases de datos grandes consiste en particionar la base en dos partes y utilizar una parte para entrenamiento y la otra para prueba. Este método fue el método utilizado para obtener los primeros resultados. El método anterior presenta limitaciones dado a que introduce sesgos debidos a la evaluación de error en la partición particular elegida. Para evitar esto se puede recurrir al método denominado leave-k-out que permite realizar particiones múltiples de los datos y luego estimar el error en base al promedio sobre todas estas particiones. Este método fue el utilizado en los experimentos más recientes.

4 Conclusiones

Debemos destacar nuevamente que el módulo de procesamiento y el modelo acústico no son tan independientes como se espera. En el diseño de los sistemas actuales del "estado del arte", se han ajustado ambos sucesivamente en una suerte de optimización combinada. Esto hace difícil su modificación por separado si sólo se espera disminuir el error. En la búsqueda de nuevos paradigmas es importante conseguir un enfoque coherente, unificado y jerárquico para la solución integral del problema del ASR en todos sus niveles y dimensiones. Esta ha sido una de las razones para el éxito alcanzado por los HMM. Se puede decir que durante la presente investigación se han explorado muchas alternativas tratando de aplicar las ventajas de las técnicas más novedosas para intentar solucionar el complejo problema del ASR. Además se han explorado otras áreas relacionadas con el habla con resultados muy prometedores, como

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. L. Rufiner & D. H. Milone; "Sistema de reconocimiento automático del habla"
Revista Ciencia, Docencia y Tecnología. Vol. XV, No. 28, pp. 149-178, 2004.

por ejemplo: identificación del hablante (Torres y Rufiner, 2001), detección automática de patologías del aparato fonador (Martínez y Rufiner, 2001), estudios de complejidad en señales de voz (Torres y cols., 2003b), desarrollo de herramientas para análisis de voz (Rufiner y cols., 1997), sistemas de rehabilitación de niños sordos (Martinez, 2001) y técnicas de mejoramiento para prótesis auditivas (Tochetto y cols., 2003). Como conclusión final creemos que se han realizado numerosos aportes al ASR, siempre en la búsqueda de "nuevas" ideas, tratando de volver a pensar algunas abandonadas prematuramente y de no perder de vista la forma en que los humanos solucionamos este problema.

5. Bibliografía

ARONSON L., RUFINER L., FURMANSKY H., ESTIENE P. "Características acústicas de las vocales del español rioplatense." Fonoaudiológica, 46(2), 2000:12-20.

DELLER J., PROAKIS J., HANSEN J. Discrete Time Processing of Speech Signals. Macmillan Publishing, New York, 1993.

GAMERO L. G., RUFINER H. L. "Paquetes de onditas evolutivas para clasificación de señales." Anales del Ier Congreso Latinoamericano de Ingeniería Biomédica, 1, 1998:784-787.

GAMERO L. G., PLASTINO A., TORRES M. E. "Wavelet analysis and nonlinear dynamics in a non extensive setting." Physica A, 246, 1997:487-509.

GODDARD J. C., MARTINEZ F. M., MARTINEZ A. E., CORNEJO J. M., RUFINER H. L., ACEVEDO R. C. "Redes neuronales y Árboles de decisión: Un enfoque híbrido." Memorias del Simposium Internacional de Computación organizado por el Instituto Politécnico Nacional, 1995.

LOPEZ-COZAR R., MILONE D. H. "A new technique based on augmented language models to improve the performance of spoken dialogue systems." EuroSpeech 2001, 2001:741-744.

LOPEZ-COZAR R., RUBIO A. J., BENITEZ M., MILONE D. H. "Restricciones de funcionamiento en tiempo real de un sistema automático de dialogo." Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, 26, 2000:169-174.

MALLAT S. G. A Wavelet Tour of signal Processing. Academic Press, segunda edición, 1999.

MARTINEZ C. E., RUFINER H. L., "Acoustic Analysis of Speech for Detection of Laryngeal Pathologies", Proceedings of the Chicago 2000 World Congress IEEE EMBS, Paper No. TH-Aa325-07, Chicago, July 2000.

MARTINEZ S., "Desarrollo de herramientas computacionales para la detección del nivel de audición y apoyo en el aprendizaje del habla para niños sordos e hipoacúsicos", Tesina de grado, Bioingeniería, FIUNER, 2001.

MILONE D. H. "Reconocimiento automático del habla con redes neuronales artificiales." 2001. Inédito.

MILONE D. H., MERELO J. J. "Evolutionary algorithm for speech segmentation." 2002 IEEE World Congress on Computational Intelligence, Hilton Hawaiian Village Hotel, Honolulu, 2002. Paper No. 7270.

MILONE D. H., RUBIO A. J. "Including prosodic cues in asr systems." Proceedings 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001), Orlando, Julio 2001. Paper No. IS0051403.

MILONE D. H., RUBIO A. J. "Prosodic and accentual information for automatic speech recognition." IEEE Trans. on Speech and Audio Processing, 11(4), 2003: 321-333.

MILONE D. H., RUBIO A. J., LOPOEZ-COZAR R. "Modelos de lenguaje variantes en el tiempo." Memorias del XXIV Congreso Nacional de Ingeniería Biomédica, Oaxtepec, México, 10-13 de Octubre 2001.

MILONE D. H., SAEZ J. C., SIMON G., RUFINER H. L. "Self-organizing neural tree networks." Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Hong Kong, 3, 1998a:1348-1351.

MILONE D. H., SAEZ J. C., SIMON G., RUFINER H. L. "Árboles de redes neuronales autoorganizativas." Revista Mexicana de Ingeniería Biomédica, 19(4), 1998b:13-26.

RABINER L., JUANG B. H. Fundamentals of Speech Recognition. Signal Processing Series. Prentice-Hall, 1993.

RUFINER H. L., GODDARD J., MARTINEZ A. E., MARTINEZ F. M. "Basis pursuit applied to speech signals." Proceedings 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001), Orlando, July 2001:517-520.

RUFINER H. L. "Comparación entre análisis onditas y Fourier aplicados al reconocimiento automático del habla." Master's thesis, Universidad Autónoma Metropolitana, Diciembre 1996.

RUFINER H. L. "Modelización biológica, redes neuronales y HMM´s aplicados al reconocimiento automático del habla." Informe de avance beca de investigación Conicet, CONICET, 1994.

RUFINER H. L., CORNEJO J. M., CADENA M., HERRERA E. "Laboratorio de voz." Anales del VIII Congreso de la Asociación Mexicana de Audiología, Foniatría y Comunicación Humana, Veracruz, México, 1997.

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. L. Rufiner & D. H. Milone; "Sistema de reconocimiento automático del habla"
Revista Ciencia, Docencia y Tecnología. Vol. XV, No. 28, pp. 149-178, 2004.

RUFINER H. L., MARTINEZ C., TORRES H.M. "Clasificación de fonemas mediante paquetes de onditas orientadas perceptualmente y una red neuronal por fonema." Anales de las "VIII Jornadas de Jóvenes Investigadores Grupo Montevideo", Brasil, Septiembre 2000.

RUFINER H. L., MARTINEZ C., TORRES H.M. "Comparación entre filtrado óptimo probabilístico y filtrado no lineal mediante redes neuronales para limpieza de habla continua con ruido." Anales de las "VIII Jornadas de Jóvenes Investigadores Grupo Montevideo", Brasil, 2000.

RUFINER H. L., ROCHA L. F., GODDARD J. "Denoising of speech using sparse representations." Proc. of the International Conference on Acoustic, Speech & Signal Processing, 2002.

RUFINER H. L., TORRES M. E., GAMERO L . E., MILONE D. H. "Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition" Physica A, 332, February 2004:496-508.

SIGURA A. "Sistema de preprocesamiento basado en paquetes de onditas y algoritmos genéticos para reconocimiento automático del habla." Tesis de Maestría, FIUNER-IPSJAE (Cuba), 2001.

TOCHETTO D., MILONE D. H., RUFINER H. L., ARONSON L. "Robust CIS Strategy for Cochlear Implants" World Congress on Medical Physics and Biomedical Engineering, Sydney, Australia, 2003:2593-2595.

TORRES H.M., RUFINER H. L. "Clasificación de fonemas mediante paquetes de onditas orientadas perceptualmente." Anales del Ier Congreso Latinoamericano de Ingeniería Biomédica, Mazatlán 98, México, 1, 1998:163-166.

TORRES H.M., RUFINER H. L., "Automatic Speaker Identification by means of Mel Cepstrum, Wavelets and Wavelets Packets", Proceedings of the Chicago 2000 World Congress IEEE EMBS, Paper No. TU--E201--02, Chicago, July 2000.

TORRES M. E., GAMERO L. G., D'ATTELLIS C. E. "A multiresolution entropy approach to detect epileptiform activity in the EEG." IEEE Workshop On Nonlinear Signal And Image Processing, II, 1995:791-794.

TORRES M. E., GAMERO, L. E., RUFINER L. H., MARTINEZ C. E., MILONE D. H., SCHLOTTHAUER G. "Segmentación automática de señales de voz mediante el análisis de cambios en la entropía multiresolución continua", XIV Congreso Argentino de Bioingeniería, III Jornadas de Ingeniería Clínica (SABI2003), Córdoba, Argentina, Artículo No. 125, 2003a.

TORRES M. E., GAMERO, L. E., RUFINER L. H., MARTINEZ C. E., MILONE D. H., "Study of Complexity in Normal and Pathological Speech Signals," 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Cancun, Mexico, 2003b:2339-2342.

VARILE G., ZAMPOLLI A. (Eds.) Survey of the State of the Art in Human Language Technology. National Science Foundation, Directorate XIIIE of the Commission of the European Communities, Center for Spoken Language Understanding, Oregon Graduate Institute, Noviembre 1995.

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. L. Rufiner & D. H. Milone; "Sistema de reconocimiento automático del habla"
Revista Ciencia, Docencia y Tecnología. Vol. XV, No. 28, pp. 149-178, 2004.