Clasificación de Fonemas mediante Paquetes de Onditas orientadas Perceptualmente y una Red Neuronal por Fonema

H. Torres(*), C. Martinez(*), H. L. Rufiner(*)

(*)UNER, Laboratorio de Cibernética, Ruta 11 Km 10, 3100 Paraná, Entre Ríos, Argentina,
hmtorres@fi.uner.edu.ar, cesarmart@amet.com.ar, lrufiner@alpha.arcride.edu.ar

Introducción

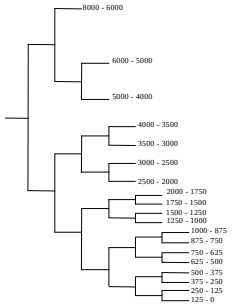
- En poco más de diez años de existencia, el área de las onditas (en inglés *wavelets*) ha llegado a ser de suma importancia para el procesamiento de señales.
- Esto se debe en gran parte a su manera natural de tratar a las señales no-estacionarias.
- En trabajos anteriores se ha utilizado la transformada Paquetes de Onditas como preprocesamiento en un sistema de Reconocimiento Automático del Habla (RAH), donde el clasificador consistía en una Red Neuronal Artificial con Retardos Temporales (TDNN, del inglés *Time Delay Neural Network*).
- En estos trabajos se exploraba cómo se comportaban distintos paquetes de onditas orientados perceptualmente según la escala de Mel, tomando el logaritmo de la energía de cada banda como patrón de entrada para el clasificador, el cual consistía en una sola gran TDNN.
- En este trabajo se presenta una nueva estructura de TDNN modular, la cual consiste en una red para cada uno de los fonemas a clasificar.
- Esto provee un significante aumento en los porcentajes de reconocimiento del sistema.
- Para realizar los experimentos se utilizó la base de datos TIMIT, ampliamente difundida en el medio, sobre los fonemas "*b*", "*d*", "*eh*", "*jh*" y "*ih*", los cuales han sido utilizados en experimentos anteriores.

Acerca de los datos

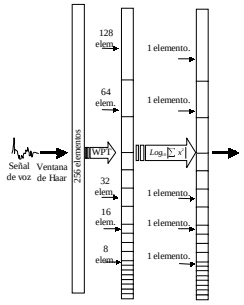
- Las señales de voz para los experimentos fueron obtenidas del corpus de voz continua TIMIT.
- Esta base de datos ha sido confeccionada en forma conjunta por Texas Instruments (TI) y el Massachusetts Institute of Technology (MIT).
- Es una de las bases multi-hablante más empleadas en el ámbito del Reconocimiento Automático del Habla (RAH) de discurso continuo por ser la más grande, completa y mejor documentada de su tipo.
- Esta base o corpus posee una gran cantidad de fonemas en diversos ambientes y pronunciados por más de 600 hablantes diferentes.
- El corpus TIMIT incluye la señal de voz correspondiente a cada oración hablada, así como también las transcripciones ortográficas, fonéticas y de palabras alineadas temporalmente.

Procesamiento de la señal de voz

- El procesamiento elegido fue la **Transformada Wavelet Packet**, integrando por bandas.



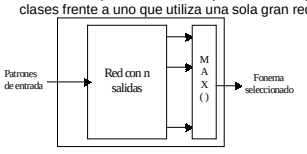
Base elegida para realizar los experimentos. Los números indican la banda de paso de cada filtro, en Hz.



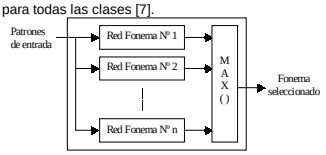
Integración por escalas de los coeficientes de la WPT.

Redes neuronales artificiales

- Una Redes Neuronales Artificiales RNA es un sistema de procesamiento de información o señales compuesto por un gran número de elementos simples de procesamiento, llamados neuronas artificiales o simplemente nodos.
- Dichos nodos están interconectados por uniones directas llamadas conexiones y cooperan para realizar procesamiento en paralelo con el objetivo de resolver una tarea computacional determinada [5].
- Dado que los patrones a clasificar son dinámicos, se hace necesario el uso de redes neuronales con retardos temporales (RNRT) [6].
- Un sistema que utiliza una red por clase es superior en robustez ante el aumento en el número de clases frente a uno que utiliza una sola gran red para todas las clases [7].



Estructura del clasificador tradicional.



Estructura de clasificador propuesto

- Las redes fueron entrenadas usando el algoritmo Backpropagation.
- Los entrenamientos se detuvieron en el pico de generalización.
- Se utilizó la función sigmoidea como función de activación no lineal.

Resultados

- En las tablas 1 y 2 se detallan los resultados obtenidos con la ondita Splines y Daubechies, respectivamente. Los mismos se obtuvieron con una TDNN, con una estructura de 19 neuronas en la capa de entrada, más 19 de retardos, 32 neuronas en la capa oculta, y una neurona de salida (lo cual abreviamos 19+19/32/1). Las redes se entrenaron con un coeficiente de aprendizaje y de momento de 0.1.
- Los mejores resultados se presentan en las tabla 3, los cuales se obtuvieron con una estructura 19+19/15/1 y la ondita Daubechies.

Tabla 1: Resultados para la ondita Splines con los patrones de entrenamiento y prueba.

Fonema	B	D	EH	IH	JH	Promedio
Entrenamiento	82.05 %	83.01 %	64.29 %	80.64 %	98.30 %	81.66%
Prueba	76.82 %	79.27 %	62.07 %	76.12 %	94.77 %	77.81%

Tabla 2: Resultados para la ondita Daubechies con los patrones de entrenamiento y prueba.

Fonema	B	D	EH	IH	JH	Promedio
Entrenamiento	83.96 %	95.41 %	85.07 %	77.79 %	100 %	88.45%
Prueba	82.12 %	76.83 %	82.73 %	76.77 %	97.67 %	83.22%

Tabla 3: Resultados para la ondita Daubechies con los patrones de entrenamiento y prueba.

Fonema	B	D	EH	IH	JH	Promedio
Entrenamiento	89.38 %	92.06 %	87.56 %	80.19 %	99.72 %	89.79%
Prueba	84.11 %	80.89 %	87.27 %	78.60 %	98.26 %	85.83%

Conclusiones y trabajos futuros

- En el presente trabajo se presenta una estructura de red neuronal modular para el reconocimiento de fonemas. Se utilizó la Transformada Paquetes de Onditas como preprocesamiento de la señal de voz. Para cada uno de los fonemas a clasificar se entrenó una red neuronal con retardos temporales.
- El sistema así conformado obtiene aumentos relativos en reconocimiento del orden del 10% respecto a un sistema con un clasificador constituído por una única red neuronal.
- Esta estructura modular presenta varias ventajas con respecto a los clasificadores utilizados en trabajos anteriores: si utilizamos una red para todos los fonemas, el procesamiento de la señal de voz es necesariamente el mismo para cada uno de los fonemas; mientras que si utilizamos una red por fonema, podemos utilizar el procesamiento de la señal de voz más conveniente a cada uno de los fonemas.
- Por lo anteriormente dicho, se debería encontrar el procesamiento óptimo para cada uno de los fonemas, para lo cual se podría implementar un algoritmo genético que incluya varios tipos distintos de procesamiento, por ejemplo: Transformada de Fourier, Coeficientes Cepstrales en escala de Mel, WT y WPT, variando a su vez las estructuras de los procesamientos, como por ejemplo: la ondita, el árbol de filtros, las posibles integraciones, etc. Para ello, se podría medir algún tipo de distancia entre clases, incluyendo una medida de la dispersión.
- En los experimentos realizados se utilizó la misma estructura para cada fonema, pero la estructura óptima de la red no necesariamente debe ser la misma para cada fonema.

Referencias

[1] Rioul, O., Vetterli, M., "Wavelets and Signal Processing", IEEE Magazine on Signal Processing, pp. 14-38, October 1991

[2] H. L. Rufiner, H. M. Torres, "Clasificación de Fonemas Mediante Paquetes de Onditas Orientadas Perceptualmen-te", Anales del 1º Congreso Latinoamericano de Ingeniería Biomédica, Mazatlán, Sinaloa, México, 1998.

[3] Garofolo. Lamel. Fisher. Fiscus. Pallett. Dahlgren. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation. National Institute of Standards and Technology. February 1993.

[4] M.A. Cody, *"The Wavelet Packet Transform : Extending the Wavelet Transform"*, Dr. Dobb's Journal, April 1994.

[5] Mohamad H. Hassoun. Fundamentals of Artificial Neural Networks. The MIT Press. 1995.

[6] J.L. Elman. "Finding structure in time". Cognitive Science 14 (1990) 179-211.

[7] H. Torres, H. L. Rufiner, "Identificación Automática del Hablante mediante Redes Neuronales", Anales del XII Congreso Argentino de Bioingeniería, Buenos Aires, Argentina, 2-4 Junio de 1999

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. L. Rufiner, C. E. Martínez & H. M. Torres; "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. Sep, 2000.