REALZAMIENTO DE LA VOZ UTILIZANDO REDES NEURONALES ARTIFICIALES

Andrés Renaud- andresrenaud@gmail.com

Trabajo presentado como monografía del curso *"Introducción a las Redes Neuronales Artificiales"*, (2007). Profesores: Dr. Andrés Pomi, Dr. Eduardo Misraji. Facultad de Ciencias, Universidad de la República, Uruguay.

Resumen:

Este trabajo presenta un modelo de RNA, que permitiría relazar la voz de distintos ambientes ruidosos. Partiendo de las características de la voz humana, se discuten los métodos existentes para cumplir esta tarea y una posible implementación en procesadores señales modernos.

1. INTRODUCCIÓN

Este trabajo es resultado de una breve investigación sobre la posibilidad construir un dispositivo que pueda extraer la señal de voz humana, con particular énfasis en las señales del habla, de un ambiente ruidoso en tiempo real utilizando redes neurales artificiales (RNA). Esto parece una idea plausible, ya que ciertamente el oído humano puede hacerlo y las RNA están inspiradas en el comportamiento las redes neurales reales.

Aquí se intenta proponer un modelo mediante el cual este objetivo podría ser logrado. Para ello se describen las características de la voz humana y los posibles tipos de ruidos que interfieren con ella. Se discuten algunos de los métodos existentes (*ver referencias*) para ayudar a realzar la voz humana de señales corruptas, ya sea por ruido acústico, interferencia, ruido de procesamiento u otras voces. Si bien ninguno** de estos procedimientos no involucran RNA, dado la versatilidad que ellas poseen, y que en principio se podría construir un conjunto de entrada-salida (E/S) representativo, es factible que ellas puedan apoyar o sustituir, total o parcialmente, estos métodos. El auge de las telecomunicaciones y la electrónica invitan a profundizar en esta dirección, ya que generan tanto las aplicaciones como los recursos tecnológicos para hacer posible que este tipo de producto sea imponible en el mercado. **médicas.

Este trabajo propone una serie de procedimientos que incluyen el acondicionamiento de señal necesario, así como algoritmos de RNA para poder cumplir esta tarea.

2. SEÑALES DE VOZ

Para poder procesar señales de voz de manera eficiente es necesario tener una idea de cómo se comportan. A continuación, se presenta una breve descripción de como se generan estas señales y las principales características que conciernen al tratamiento de señales.

2.1 Generación de la voz

La voz se genera en la laringe, el aire empujado por los pulmones, el diafragma y algunos músculos de la caja toráxica excitan las cuerdas vocales, produciendo dos tipos de sonidos: sonoros y sordos, según hagan vibrar o no las cuerdas vocales. Los sonidos sordos son una señal más difícil de tratar ya que por naturaleza presenta un ruido aleatorio que se genera en las angostas contracciones del tracto vocal [7].

Para ambos sonidos la garganta, el pecho y las cavidades de la cabeza actúan como filtros que amplifican ciertas frecuencias mientras que atenúan otras. La señal de voz resultante tiene un espectro que consiste de una frecuencia fundamental y sus harmónicos; algunos de los algoritmos, como los que mencionan más adelante, se basa en estimar esta frecuencia para poder extraer la señal de voz correctamente.

2.2 Características

La señal de voz pose ciertas características que deben ser tomadas en cuenta para poder realizar una correcta adaptación de señal; de esta forma la información relevante es resguardada y las hipótesis necesarias para aplicar los algoritmos descriptos más adelante válidos.

La figura 1 muestra un ejemplo de una señal de voz. Se dice que este tipo de señales es quasi-estacionaria debido a que sus características son bastantes estacionarias cuando son observadas a través

de ventanas de tiempo cortas (5-100ms), mientras que cuando se las mira sobre periodos de tiempo más largos (>200ms), la señal es no-estacionaria, reflejando los distintos sonidos la voz emitida [2]. Esto es muy importante, ya que al elegir el numero de muestras sobre el cual aplicar el algoritmo es deseable que las características de la señal permanezcan lo más constante posibles para así poder hacer una buena estimación de los parámetros que luego modificarán las señal.

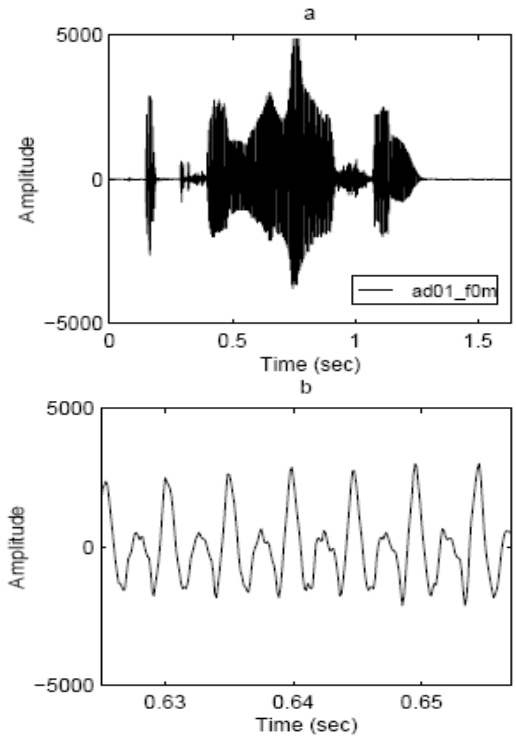

Figura 1, a: señal no-estacionaria, b: señal estacionaria; tomada de [2].

Existen otras característica de las señales de voz que generan dificultades a la hora de extraer propiedades que permitan mejorarlas desde punto de vista cualitativo, a continuación un resumen [2]:

1. **Alta redundancia:** Las señales de la voz son redundantes en corto plazo. Esta redundancia que es útil para los humanos puede traer errores y hacer ineficiente la estimación de parámetros de señal del habla, excepto en obtener una estimación promedio de las características de ciertos ruidos.

2. **Degradación de la señal:** Distintos ruidos aditivo y convolucional degradan la señal pura de la voz por diferentes razones: Ruido ambiente, características del canal de transmisión y de los micrófonos (ancho de banda limitado y variabilidades temporales), reverberación de la habitación, etc.

3. **Varibilidades temporales y en frecuencia:** Las señales de voz y el habla son producidas con diferentes velocidades y frecuencias. El sistema implementado debe ser lo suficientemente robusto para comportarse bien ante estas diferencias.

4. **Múltiples orígenes de la voz:** En grabaciones espontaneas múltiples voces pueden competir y solaparse en el tiempo.

Es conveniente aclarar a esta altura que este tipo de procesamiento de señal se pude dividir en dos aplicaciones bien distintas, que tienen que lidiar distintos tipos de ruido. La primera es aquella que realza la voz grabada con los ruidos *¨naturales¨* que provienes del ambiente y de las características micrófono. La segunda es en la reducción de ruido pos procesamiento de la señal, donde estas puden presentar ruidos debido a algoritmos de compresión, filtros, errores producidos por transmisiones u otros.

Este trabajo apunta principalmente a la primera clase de aplicaciones. Los puntos siguientes tratan la obtención del conjunto de E/S necesario para entrenar y validar el algoritmo y el pre-procesamiento que debe hacerle a las señales para mejor rendimiento de los sistemas a implementar.

3. CONSTRUCCIÓN DEL CONJUNTO E/S

Para entrenar correctamente los algoritmos de RNA es necesario obtener un conjunto E/S representativo del espacio de señales a las cuales se desea procesar. También es necesario obtener otro conjunto con similares características para validar el sistema.

Para la construcción del conjunto E/S se utilizan los software existentes de edición de sonido los cuales mezclar las señales limpias con señales de solo ruido y así obtener la entrada al sistema. Esto de hacerse con particular cuidado, ya que no es lo mismo mezclar la señal en un entorno de software, que un ambiente real donde se dan fenómenos de interferencia, intermodulación y resonancia. Es necesario hacer un estudio previo para ver que tan diferentes son estas señales ruidosas y como reducir esta diferencia. Con este fin se puede correr un análisis estadístico ambos grupo de datos, como estudios de correlación y distribución de la muestras.

En una primera etapa, herramientas de software como matlab permiten construir un conjunto E/S y probar los algoritmos en un ambiente más controlado.

4. PRE-PROCESAMIENTO DE SEÑAL

Es esencial realizar pre-procesamiento mínimo a las señale para que los algoritmos se comporten debidamente. En muchos de los métodos usados para realzar la voz, como substracción espectral, necesitan la hipótesis mínima de que el ruido sea relativamente

más estacionario que la voz, estadísticamente hablando. En la practica las señales de ruido permanecen estacionario por 100ms, mientras que las señales del habla lo hacen por períodos que van de los 10 a 20 ms[3]. Por esta razón además de lidiar con la no-estacionareidad de la voz en lapsos de tiempos largos, la señal de audio se divide en marcos ó "frames" que duran de 10 a 40ms[2].

Para lograr que los algoritmos sea mas eficientes y puedan ser llevados a tiempo real se debe extraer información acústica reducida y relevante, evitando así el problema de la redundancia en el cortos períodos de tiempo. El método mas usado es pasar la señal del dominio del tiempo al dominio de la frecuencia, donde la periocidad de la señal es representada por valores mayores de energía a la frecuencia correspondiente [2].

5. MÉTODOS EXISTENTES PARA EXTRAER LA SEÑAL DE VOZ

Existen varios métodos para la extracción de una señal de voz de un ambiente ruidoso. Los distintos planteos pueden ser efectivos o no dependiendo de la aplicación y los distintos tipos de ruido a los que nos enfrentamos. Van desde algoritmos simples como substracción espectral, que tienen buena performance en presencia de ruidos relativamente estacionarios, a métodos que involucran algoritmos más complejos como la separación de origen ciego (BSS, blind source separation), capaz de extraer la voz de un individuo en presencia de ruidos musicales y otras voces.

Si bien ninguno de estos trabajos([5],[6],[7]) plantea la utilización de RNA para llevar acabo su cometido, es difícil no ver una relación con las RNA ya que usan sumas ponderas de elementos cuyos pesos son modificados minimizando errores y equivalentes a funciones de activación que conllevan a funciones no lineales; sin embargo, no describen formas de aprendizaje para los algoritmos según un conjunto E/S para el entrenamiento, sino que las modificaciones se realizan en función de la entrada actual y entradas anteriores.

A continuación se hace una breve descripción de las características fundamentales de estos métodos .

5.1 Substracción Espectral

La substracción espectral es muy popular en la supresión de ruido en señales de habla. Esta técnica de realzamiento de la voz esta basado en estimación directa del espectro en el corto plazo. La señales del habla se modelan como procesos aleatorios a los cuales se les adiciona un ruido aleatorio no correlacionado,

$$x[n]=s[n]+n[n] \qquad (1)$$

donde s[n], n[n],x[n] representan señales estocásticas de la voz, ruido y voz ruidosa respectivamente. Se asume que el ruido es estacionario en corto plazo y es estimado durante marcos en las cuales no hay señal de voz. Al espectro de la señal ruidosa se le resta el espectro de ruido estimado [3],[5].

Este algoritmo es robusto, de fácil implementación y computacionalmente eficiente. Pero tiene ciertas limitaciones respecto ruidos varían rapidamente, además es necesario detectar previamente que marcos de señal posen voz o no, lo que puede no ser una tarea fácil.

En [3] y [5] proponen combinaciones de este método con filtros adaptativos, como el filtro de Wiener (ec. 2), las cuales permiten estimación continua del ruido de fondo y una sensible eliminación de los tonos musicales. También describen operaciones previas que le pueden hacer a las señales para obtener una mejor performance del sistema, así como promediar variaciones en los parámetros para obtener mejores resultados ante ruidos que varían más rápido.

Sin embargo estos métodos no pueden reducir efectivamente los ruidos no-estacionarios como otras voces o señales musicales. Los métodos de BSS ofrecen mejores resultados en la extracción de una señal particular de voz en bajo condiciones menos determinadas.

5.2 Separación de Origen Ciego. (BSS)

El método clásico para extraer una voz determinada es el Cancelador generalizado de lóbulo lateral (GSC, en *inglés*); el reconocido por ser la forma más practica y efectiva de extraer las señales de una fuente con varias señales, pero resulta insuficiente en la reducción de ruido.

El algoritmo **BSS** puede ser usado para separar una señal objetivo en presencia de ruidos no-estacionarios. Al igual que GSC, este algoritmo utilizan un arreglo de micrófonos para poder distinguir la señal objetivo. Esto se basa en que las distintas señales captadas por los micrófonos tendrán ganancia y retardo particulares dependiendo de la posición en que se encuentran las fuentes.

Hay al menos dos formas de construir un algoritmos BSS. La primera basada en el análisis de componente independiente (ICA, Idependent Component Analysis) y la segunda se basa en un enmascaramiento en el dominio tiempo-frecuencia (TF). En ejemplo de estos algoritmos son presentados en [7] y [6] respectivamente.

5.2.1 Análisis de Componente Independiente.

Este método asume que las distintas señales eque entran en la grabación son independientes y que aplicado ICA se puede reconstruir las distintas señales. Bell y Sejnowiski propusieron un método que maximiza la información (*infomax*), como algoritmo ICA [6].

El principio infomax sapara las señales idependientes utilizando la matriz de separación W. Los coeficientes de esta matriz son entrenados

minimizando la información mutua entre las componentes *y(t)=g(u(t))*, donde g es una función no lineal que se aproxima la función de densidad acumulativa de las distintas fuentes de sonidos. El minimizar la información mutua entre las componentes de *y* equivale a maximizar la entropía de *y*. Para lograr esto se modifican los coeficientes de separación según el algoritmo de gradiente estocástico ascendente [6].

En [7] se propone un algoritmo ICA que trata imitar el funcionamiento del oído humano emulando la cóclea, esta realiza un análisis espectro-gráfico del estímulo auditivo que puede ser considerado como un banco no-uniforme de filtros auto-adaptables. Siguiendo este funcionamiento el algoritmo divide la señal se divide en sub-bandas; luego, utiliza un algoritmo BSS-ICA para determinar que sub-banda posee información relevante de la fuente se señal que desea realzar. Se utiliza un arreglo de al menos dos micrófonos. La señal realzada será aquella que se encuentre más cerca del punto medio del arreglo.

Este método produce buenos resultados; sin embargo, resulta demasiado lento para hacer una implementación que funcione en tiempo real [6].

5.2.2 Máscaras T-F

Este método se basa en la hipótesis de la señales de distintas fuentes de ruido e interferencia son casi ortogonales disjuntas cuando observadas en cortos períodos de tiempo, W-DO (Widowed-Disjoint Orthogonal), en el dominio tiempo-frecuencia. De esta forma se logra una implementación en tiempo de real más fácilmente y se comporta mejor que ICA-BSS bajo condiciones poco determinadas [6].

En [6] también se utiliza un arreglo de micrófonos para extraer las características direccionales de la señal, de manera de realzar aquella que se encuentra enfrente del arreglo. Luego utiliza enmascaramientos en el espacio tiempo-frecuencia para obtener la señal objetivo.

6. MODELO PROPUESTO.

Debido a la gran versatilidad de los algoritmos implentados por las RNA, ellos podrían asistir o sustituir en parte a la mayoría, sino a todos, los métodos anteriormente mencionados. Podría elegir los lo parámetros óptimos del filtro de Wiener en las substracción espectral. Serian ser útiles para decidir que marco de señal o sub-banda contiene información de la señal que se desea extraer. Sin embargo un modelo más osado, que es puramente implementado por redes neuronales se presenta en [4]; y en él se basa el modelo propuesto.

En este trabajo el autor muestra una configuración de RNA que actúa como filtros de frecuencias. Si bien la aplicación que presenta, es simplemente filtro pasa bajos, fácilmente implementable de forma efectiva por métodos ya conocidos, es interesante investigar como se comportaría ante una tarea un tanto más compleja como realzar la voz.

6.1 Arquitectura de la RNA.

Esta red esta compuesta por neuronas que se asemejan al perceptrón multicapa, con la diferencia de la capa oculta, no es tal.

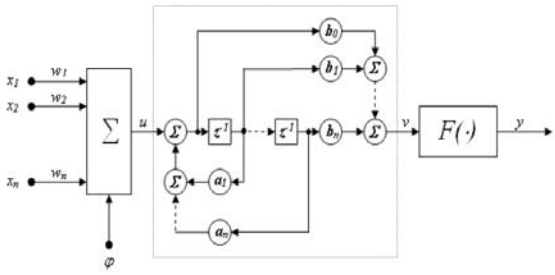

Figura 2 Diagrama de bloques la neurona dinámica

En la figura se observa lo que el autor de [4] llama ***neurona dinámica***. La capa de entrada por una capa de entrada la hace una suma ponderada de los vectores de entrada.

$$u(k) - \sum_{i=1}^{n} w_i x_i(k) + \varphi \quad (3)$$

La salida de esta capa es luego promediado ponderadamente con los valores anteriores de la entrada u y salida v para formar la salida en el instante actual. La entrada y salida de esta capa se relacionan mediante expresión matemática que representa los filtros IIR.

$$v(k) = a_0\, u(k) + a_1\, u(k-1) + .. + a_n\, u(k-n) + b_1\, v(k-1) + .. + b_m\, y(k-m) \quad (4)$$

Donde los coeficientes a_i y b_i son los pesos hacia atrás (*feedback*) y hacia delante (*feed forward*), respectivamente. Finalmente la salida de la capa dinámica pasa por la función de activación no lineal, donde γ es el parámetro que fija la pendiente de la función de activación.

$$y(k) = \frac{1}{1+e^{-\gamma v(k)}} \quad (5)$$

Lo siguiente es definir que tipo de par entrada-salida usará el algoritmo. Una opción es alimentar la red con parámetros de previamente estimados, como derivadas, máximo, periodicidad en dominio de tiempo y frecuencias y otros que ayuden determinar la percepción acústica del sonido. Pero esto requiere un pre-tratamiento complejo de la señal y una forma para

modificar la señal acústica (por ejemplo un filtro adaptativo) a partir de los parámetros obtenidos.

Otra opción mucha más simple de implementar, pero que podría llevar a una convergencia más esquiva, dado que no se tiene ningún control sobre los datos que ingresan al sistema, es uno en el cual la entrada $(x_1(k),\ldots ,x_n(k))$, sea la muestra digital de la señal analógica en el instante k, la cual entra al bloque de la figura 2 para formar la salida y_i, y de esta forma con n bloques adyacentes obtener la salida $(y_1(k),\ldots ,y_n(k))$. Los pesos de la capas de entrada fácilmente podría ignorar errores introducidos por el muestreo de la señal u otros fenómenos que se representen en la variabilidad de algunos que bits que no deberían afectar la salida. Los coeficientes de la capa dinámica deberían activar las salidas en función de las características temporales de la señal.

6.2 Reglas de aprendizaje

Otra parte decisiva para lograr que el algoritmo funcione es elegir formas adecuadas para modificar los pesos según la salida obtenida y la salida esperada. En [4] el autor propone utilizar la función de error definida en la siguiente ecuación donde y_d es la salida esperada.

$$J(k) = \frac{1}{2}E\left\{\left(y_{d(k)} - y(k)\right)^2\right\}$$

(6)

Luego ultiliza el algoritmo de gradiente decendiente para encontrar los parámetros de la RNA que minimizan el error. Otros algoritmos de aprendizaje como backpropagation podrían utilizarse para observar la performance de la red.

6.3 Otras modificaciones.

Una vez que se logre verificar que el algorito propuesto funciona para reducir ruidos no estacionarios en señales de voz, se podria utilizar un arreglo de micrófonos como los mencionados anteriormente para realzar la voz en la dirección deseada, eliminado interferencias provenientes de otras voces.

6.4 Implementación hardware

Gracias al gran avance que la electrónica ha logrado en los últimos años, podemos contar una plataforma lo suficientemente rápida para poder correr el algoritmo en tiempo real. Si bien el algoritmo antes propuesto parece a primera vista bastante pesado en cuanto a requerimientos informáticos, el desarrollo de microcontroladores específicamente diseñados para el procesamiento de señales digitales hace la propuesta un tanto más probable y serían la primera opción para intentar desarrollar un equipo que cumpla con el objetivo propuesto.

Existen dos tipos de equipos ideales para este tipo de implementación, el DSP (Digital Singnal Processors) y el FPGA (Field Programmable Gate Array). Los DSP son microcontroladores diseñados para el tratamiento de señales en tiempo real se caracterizan por un menor costo, progración relativamente fácil (lenguage C), y además posee una mayor cantidad de herramientas para el diseño y depuración del programa.

Los FPGA, sin embargo, están en auge en la linea de procesadores de señales, facilmente pueden superar en velocidad la los DSP gracias a que pueden procesar muchos datos en paralelo; además, pueden ser una solución menos costosa si fuera necesario utilizar más de un DSP en paralelo para lograr el mismo rendimiento. Posee la deventaja de no tener tantas herramientas que ayudan al pogramador, como en los DSP, lo que conyeva aun mayor inverción de tiempo para hacer que el algoritmo funcione.

7. CONCLUSIONES

En este trabajo se ha presentado un modelo de red neuronal, formado por neuronas multicapas, que en principio podría lograr el objetivo de extraer la voz de una señal ambiente corrupta por distintos ruidos y otras voces. También se observó que las RNA pueden complementar los métodos existentes para realzar la voz ya muchos de ellos involucran problemas de clasificación y las RNA se destacan en esta función.

El desarrollo de equipos especialmente diseñados para procesar señales hace plausible una implementación en tiempo real del algoritmo propuesto. Pero sobre todo este trabajo incita seguir estudiando y comprobar la utilidad de las redes neuronales artificiales.

8. REFERENCIAS

[1] Novel Approach to Acustical Voice Analysis Using Artificial Neural Networks. Rainer Schonweiler, Markus Hess, Peter Wuebbelt y Martin Ptok. JARO (2000)

[2] Hidden Markov Model and Artificial Neural Networks for Speech and Speaker Recognition. Jean Hennebert. Tesis de posgrado de la Escuela Politécnica de Lausanne (1998).

[3] Adaptive, Acustic Noise Supression for Speech Enhacement. Phil Whitehead, David Anderson, Mark Clements. Giorgia Institute of Technology, Atlanta (2003).

[4] Artificial Neural Network as Frecuency Filters. Andrzej Zak. Naval University of Gdynia, Poland.

[5] Implementation of Spectral Substraction Modified by Wiener Filtering on Fixed Point DSP. Vratislav Davidek. Czeck Technical University.

[6] On-line Speech Enhancement by Time-Frecuency Masking Under Prior Knowledge of Source Location. Min Ah Kang, Sangbae Jeong, Minsoo Hahn. PWASET (2007).

[7] Speech Enhacement Using Adaptive Filters and Independent Component Analysis Approach. Tomasz Rutkowski, Andrzej Cichocki, Allan Kardec Barros. Brain Science Institute RAIKEN.