PREDICCION NO LINEAL DE VOZ BASADA EN RED NEURONAL

Marcos Faúndez Zanuy, Enric Monte Moreno
Departamento de Teoría de la Señal y Comunicaciones
Universidad Politécnica de Catalunya
Campus Nord C/Gran Capitán s/n, Edificio D5, 08034 BCN
e-mail: marcos@gps.tsc.upc.es

ABSTRACT

Speech applications usually require the computation of a linear prediction model for the vocal tract. This model has been successfully applied during the last thirty years, but it has some drawbacks. Mainly, it is unable to model the nonlinearities involved in the speech production mechanism, and only one parameter can be fixed: the analysis order. With nonlinear models, the speech signal is better fit, and there is more flexibility to adapt the model to the application.

1. INTRODUCCION

Una gran parte de las aplicaciones relacionadas con el tratamiento del habla, están basadas en al análisis de predicción lineal LPC [1]. Este modelo es análogo a representar el tracto vocal mediante un tubo acústico sin pérdidas excitado por un extremo, con diversos tramos de secciones apropiadas. Este modelo resulta inexacto, puesto que el tracto vocal presenta pérdidas, en las fricativas la excitación no se produce al inicio del tracto, y en los sonidos nasálicos existe una cavidad adicional que no se tiene en cuenta. Por otra parte, se han observado no linealidades en la señal de voz [2].

Mediante modelado no lineal de series temporales, puede conseguirse una mayor adecuación del modelo a la realidad. Como aplicaciones potenciales cabe destacar todos aquellos campos en los que tradicionalmente se ha venido utilizando el modelado LPC (análisis, reconocimiento de locutor y del habla, codificación, síntesis, etc.). Como aspectos negativos, cabe destacar un mayor coste computacional, y la dificultad intrínseca que conlleva el análisis de sistemas no lineales [3].

Este artículo consta de las siguientes partes: el apartado 2 resume las principales características del análisis de predicción lineal. El apartado 3 trata las alternativas más comunes para realizar predicción no lineal y compara las principales ventajas e inconvenientes frente al análisis LPC. El apartado 4 evalúa los diferentes parámetros existentes al trabajar con redes neuronales, y finalmente se presentan las conclusiones más relevantes.

2. PREDICCION LINEAL

En el análisis de series temporales, el modelado de señal más utilizado, consiste en expresar la señal y en el instante **n** (y[n]) como una combinación lineal de entradas y salidas en instantes anteriores y la entrada actual:

$$y[n] = \sum_{K=1}^{p} a_K y[n-K] + G\sum_{l=0}^{q} b_l x[n-l] \qquad (1)$$

Los parámetros del modelo son: a_K, $1 \le K \le p$; b_l, $1 \le l \le q$; y G. Pueden calcularse con la formulación dada en la referencia [1].

Estos parámetros permiten predecir la señal y[n] a partir de combinaciones de salidas y entradas anteriores. De aquí el nombre de predicción lineal. En nuestro caso nos limitaremos a sistemas que realizan la predicción únicamente a partir de las entradas anteriores.

Para evaluar la bondad del modelo suele utilizarse la ganancia de predicción, definida como una relación de potencias:

$$G_p\ [dB] = 10\log\frac{E[y^2[n]]}{E[e^2[n]]} \qquad (2)$$

donde e es el error residual de predicción (diferencia entre la señal original y su predicción):

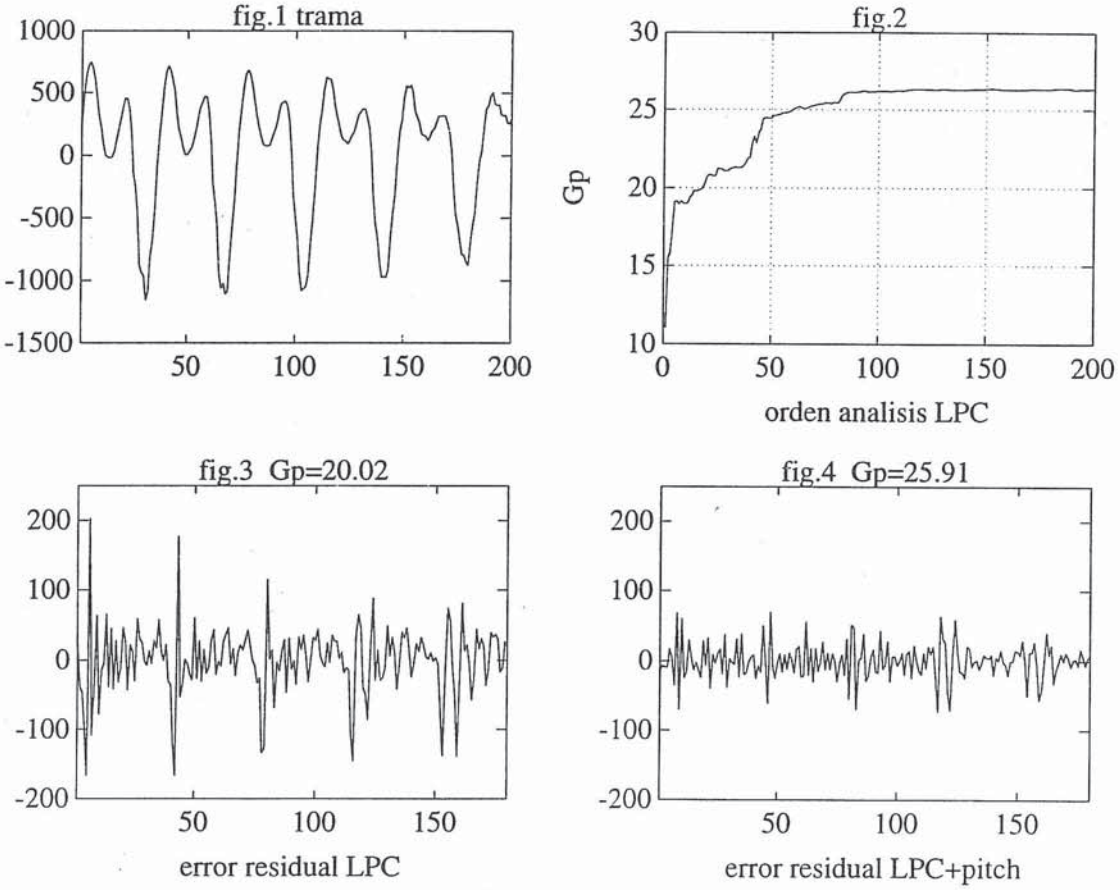
$$e[n] = y[n] - \hat{y}[n] \qquad (3)$$

Fijado el modelo nos queda como parámetro a escoger el orden q del análisis. Para un fragmento (trama) de señal de voz (fig. 1) se obtienen las ganancias de predicción (en función de q) de la figura 2. Suele utilizarse un orden de predicción q=10, obteniéndose en las partes sonoras un error residual de predicción (en el punto A de la figura 5) semejante al de la figura 3. En el error residual destacan unos picos cuya separación coincide con

el periodo de pitch de la trama. Este resultado es característico de los predictores lineales a corto plazo. Para eliminarlo puede añadirse en cascada un predictor a largo plazo $P_L(z)$ consistente en utilizar muestras retardadas M instantes de tiempo, donde M es el periodo de pitch de la trama, expresado en muestras.

$$P_L(z) = \sum_{j=-K}^{K} b(K) z^{-M-j} \tag{4}$$

Aplicando predicción de pitch se obtiene un error de predicción en el punto B de la figura 5 que está representado en la fig. 4. Se observa que los picos de error han disminuido y la ganancia de predicción ha aumentado. Sin embargo, este predictor es extremadamente sensible a errores en la estimación y/o transmisión del valor de pitch M. Por otra parte, para tamaños de trama mayores o tramas menos periódicas, el espaciamiento entre picos no es constante y el predictor P_L empeora la ganancia de predicción de todo el sistema.



La fig. 1 muestra una trama de voz sonora. La fig. 2 representa la ganancia de predicción (Gp) en función del orden del análisis LPC. La fig. 3 es el error residual de predicción del análisis LPC 10. La fig. 4 presenta el error de predicción después de aplicar análisis LPC 10 y predicción de pitch.

El uso combinado de los dos predictores es análogo a la predicción bidimensional en tratamiento de imagen, en el que se utilizan pixels de distintas líneas. La periodicidad de pitch de la voz se corresponde con el tiempo de línea. En el caso de imagen, a diferencia de la voz, existe la ventaja de que el tiempo de línea es una magnitud constante y conocida a priori.

3. PREDICCION NO LINEAL

Existen diversas alternativas para realizar predicción no lineal de una serie temporal. Las más utilizadas son las series de volterra [5] y las redes neuronales [3]. Las series de Volterra requieren un número total de coeficientes mayor a las redes neuronales, y presentan el problema de inestabilidad potencial en el filtro inverso (al realizar la síntesis a partir del error residual de predicción). Las redes neuronales no presentan problemas de estabilidad, debido a la saturación impuesta por la no linealidad , que impide el crecimiento indefinido de la salida (se trata de un sistema BIBO : Bounded Input Bounded Output). Sin embargo, no es posible calcular los coeficientes de forma analítica, y requieren costosos algoritmos iterativos de minimización del error de

predicción.

Para obviar el problema de inestabilidad, y dado que se obtienen mejores resultados con un número menor de parámetros que las series de Volterra ([3]) hemos escogido las redes neuronales.

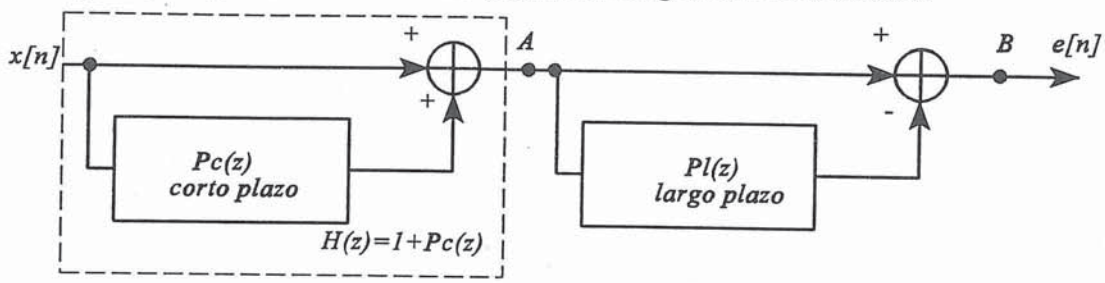


Figura 5. Esquema de filtrado para obtener el error residual de predicción.

3.1 Ventajas de la predicción no lineal frente al análisis LPC

Cabe destacar:

- Posibilidad de controlar el error residual en un margen amplio: para ello basta con variar el número de epochs sin necesidad de cambiar el tamaño de la ventana, orden del análisis, estructura, etc. Por otra parte puede eliminarse la periodicidad de pitch en el error de predicción sin necesidad de incorporar predicción a largo plazo ni transmitir/medir el pitch.
- Posibilidad de integrar información: el análisis LPC se basa únicamente en la señal de entrada. Con redes neuronales existe la posibilidad de entrar información adicional, como por ejemplo la energía de la ventana, tipo de trama (sonora, sorda, transición, etc.), cruces por cero, etc.

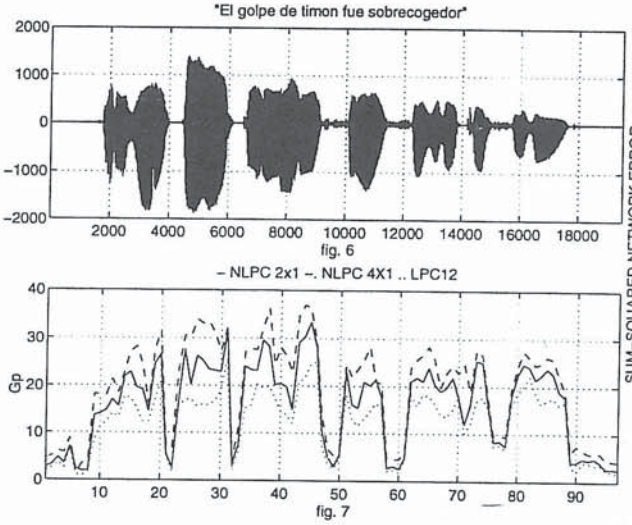
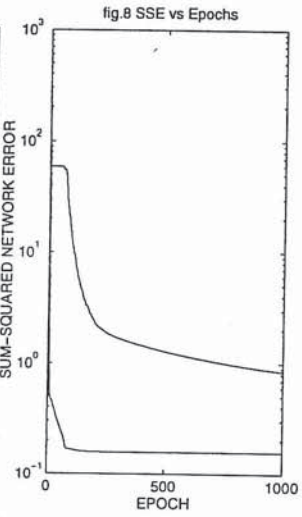


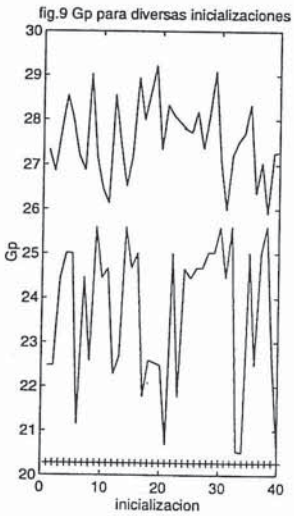
fig. 6

fig. 7

La fig. 6 representa una señal de voz, y la figura 7 las ganancias de predicción de sus correspondientes tramas utilizando 3 predictores distintos: la curva de valores mayores está obtenida con un MLP de 4x1 neuronas, la segunda con un MLP de 2x1 y la punteada con análisis LPC de 12 coeficientes.



La fig. 8 representa la evolución del error en función del nº de epochs para dos algoritmos de aprendizaje: la superior es el back-propagation y la inferior el algoritmo Levenberg-Marquardt.



La fig. 9 muestra las ganancias de predicción para 40 inicializaciones distintas: la curva superior es MLP 4x1, la inferior MLP 2x1 y + es LPC-12

3.2 Inconvenientes de un predictor no lineal

Cabe destacar:

- Mayor tiempo de cálculo. En una aplicación de codificación éste problema es obviable si se usa estructura basada en un codebook de predictores, puesto que el entrenamiento de las redes sólo debe realizarse una vez.
- Dificultad intrínseca del análisis e implementación de sistemas basados en predicción no lineal al no cumplirse el principio de superposición.
- Sobreentrenamiento de la red: aumentando considerablemente el nº de epochs puede reducirse el error de

predicción, pero a costa de sobreespecializar la red en las tramas usadas en el entrenamiento, penalizando sus prestaciones con tramas distintas.

4. EXPERIMENTOS DE PREDICCION NO LINEAL USANDO UN PERCEPTRON MULTICAPA

Una vez escogida la estructura de red neuronal para realizar la predicción no lineal, deben fijarse sus parámetros. Los más relevantes son:

Algoritmo de entrenamiento: Uno de los algoritmos más populares es el BP (back propagation). Sin embargo, requerire muchas epochs y por tanto el tiempo de cálculo es considerable. Otra posibilidad es utilizar el algoritmo de L-M. (Levenberg-Marquardt), que resulta significativamente más rápido aunque requiere más memoria. La figura 8 representa la evolución del error en función del número de epochs al aprender la trama de la figura 1. La curva inferior está obtenida con el algoritmo L-M y la superior con el BP. Observar que el BP tiene una convergencia lenta, y el resultado final corresponde a un error más elevado que el obtenido con L-M (incluso con más de 1000 epochs, puesto que la pendiente cada vez es menor).

Arquitectura de la red: Hemos estudiado el perceptrón multicapa (MLP) y la red de Elman. En principio la realimentación de las capas intermedias a la entrada (red de Elman) permite una mayor flexibilidad, pero nuestros experimentos revelan una mayor dificultad para entrenarla y un resultado peor que el MLP entrenado con LM.

Tamaño de la red: Una vez escogida la estructura de la red hay que fijar el número de capas y el número de neuronas por capa. En cuanto al número de capas, hemos obtenido buenas prestaciones usando 10 muestras de entrada, 2 neuronas en la primera capa y una capa con una neurona de salida. Utilizando más capas el entrenamiento de la red resulta más complejo. Utilizando cuatro neuronas en la primera capa se obtienen unas ganancias de predicción mayores, a costa de aumentar el tiempo de entrenamiento y el número de parámetros de la red (10x4 + 4x1 pesos y 4 + 1 bias frente a 10x2 + 2x1 pesos y 2 + 1 bias). La figura 7 muestra las ganancias de predicción obtenidas con las tramas de la frase "El golpe de timón fue sobrecogedor" (fig. 6) para análsis LPC-12 (línea punteada), MLP 2x1 (contínuo) y MLP 4x1 (discontínuo). Observar que las mejoras más significativas se obtienen en las partes sonoras.

Número de epochs: El algoritmo L-M presenta la característica de una rápida convergencia, y una reducción del error inapreciable para más de 50 ó 100 epochs. Sin embargo, es extremadamente sensible a la inicialización de los pesos. La figura 9 muestra las ganancias de predicción obtenidas con inicializaciones distintas (aleatorias) al entrenar con la trama de la figura 1. La curva superior está obtenida con una red MLP 4x1 y la inferior con MLP 2x1. La línea inferior (+) es la ganancia de LPC-12.

Una conclusión importante es que resulta más adecuado realizar menos epochs y varias inicializaciones que no una única inicialización y muchos epochs. (Por ejemplo 10 inicializaciones y 50 epochs en vez de 1 inicialización y 500 epochs. El único inconveniente es que habrá que guardar dos conjuntos de pesos: el actual y el que haya proporcionado un mejor resultado hasta el momento actual).

5. CONCLUSIONES

Pese a la dificultad intrínseca que conlleva el análisis de sistemas no lineales y su mayor coste computacional, creemos que la mejora de prestaciones obtenida respecto al análisis clásico de predicción lineal justifica el interés de su estudio y la aplicación en los diversos campos de tratamiento del habla como sustituto o soporte al análisis LPC. En cuanto a su implementación práctica, las redes neuronales constituyen una alternativa adecuada dada la gran flexibilidad que presentan. Sin embargo, quedan todavía muchos problemas por resolver y/o comprender en el procesado no lineal.

6. REFERENCIAS

[1] J. Makhoul. "Linear prediction: a tutorial review". Proceedings of the IEEE vol. 63 pp.561-580, Abril 1975.

[2] H.M. Teager "Some observations on oral air flow vocalization" IEEE trans. ASSP, vol.82 pp.559-601, Octubre 1980

[3] J. Thyssen, H. Nielsen y S.D. Hansen "Non-linear short-term prediction in speech coding". ICASSP-94, pp.I-185 a I-188.

[4] R.P. Lippmann. "An introduction to computing with neural nets". IEEE trans. ASSP. Vol.3 nº4 pp.4-22. 1988.

[5] I.Pitas y A.N. Venetsanopoulos "Non-linear digital filters: principles and applications". Ed. Kluwer 1991.