Realce de la señal de voz usando criterios no cuadráticos de filtrado

ENRIQUE MASGRAU GOMEZ, LUIS VICENTE BORRUEL
DEPARTAMENTO DE INGENIERIA ELECTRONICA Y COMUNICACIONES
CENTRO POLITECNICO SUPERIOR. UNIVERSIDAD DE ZARAGOZA.
Correo electrónico: masgrau@posta.unizar.es

Abstract:

We study a new algorithm for speech enhancement based on the iterative Wiener filtering. In our case, we propose the use of a generalized non-quadratic cost function. In addition to the classical MSE term (quadratic term), the proposed cost function includes two signal-error cross-correlation terms. These terms work to reduce both the residual noise and the signal distortion in the enhanced speech. So, we add a L2 norm term of the filter weights in order to reduce the overall gain of the filter. We expound two solutions type to the new cost function: the classical non-causal type (ideal Wiener), working in the frequency domain; and a causal finite length in the time domain. In both cases, the consecutive iterations are carried out like the Lim algorithm such as to the filter output of each iteration is used as "noiseless" speech signal for the following one. Simulation results demonstrate the effectiveness of these algorithms.

1. Introducción

Es bien sabido que muchas aplicaciones de tratamiento de voz que trabajan muy bien en condiciones de laboratorio, sufren una alta degradación de sus prestaciones cuando trabajan en ambientes reales. Esta falta de inmunidad ante ambientes hostiles representa un grave inconveniente a la hora de la a comercial de estas tecnologías del habla, lo que ha llevado a concentrar en los últimos tiempos mucho interés y esfuerzo en el desarrollo de técnicas y algoritmos robustos. En esta comunicación se trata el problema de realzar la calidad de una señal de voz enmascarada en ruido aditivo en el caso de que solo se dispone de la señal ruidosa frente al caso, denominado como cancelación de ruido, en el que se dispone de una referencia adicional del ruido, y que resulta ser un problema mucho más sencillo. Uno de los más populares algoritmos de realce de voz es el denominado Filtrado Iterativo de Wiener, formulado originalmente por Lim y Oppenheim [1]. Este algoritmo consiste en un filtrado iterativo de Wiener de la señal ruidosa, donde el cálculo del filtro se basa en una estimación espectral del ruido, obtenida en las zonas de silencio de la señal, y en un modelado AR de la voz. Este modelado espectral de la voz es mejorado de forma continua en cada nueva iteración mediante el uso de la voz realzada obtenida en la iteración precedente. La convergencia de este algoritmo se ve perjudicada por dos problemas que influyen en la exactitud del modelado AR de la voz: la influencia del ruido residual aún presente en la señal obtenida en cada iteración y, lo que es más grave, la distorsión espectral que la señal de voz sufre en cada filtrado, que tiene un carácter creciente con el número de iteraciones. Esta distorsión espectral consiste en un estrechamiento o "picado" de los formantes de la voz que produce un sonido no natural y una apreciable pérdida de inteligibilidad de la misma. Para reducir este problema, el autor propuso en otros trabajos previos [2-4] algunas soluciones basadas en el uso de estadísticas de orden superior (HOS). En estas soluciones se explotan las propiedades de desacoplo entre el ruido, supuesto gausiano, y la señal de voz, inherentes al HOS. Los resultados obtenidos son muy buenos pero el coste computacional es muy alto. En este trabajo se propone un nuevo algoritmo basado en el uso de una función de coste generalizada para el diseño del filtro. Antes de pasar a describir el nuevo algoritmo, en la siguiente sección se realiza una breve revisión de los conceptos más característicos del método iterativo de Wiener original.

2. Método original de Filtrado Iterativo de Wiener

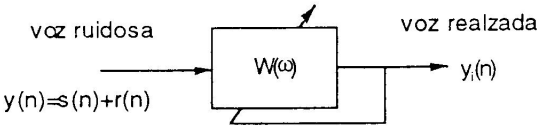
En le método original de Lim-Oppenheim [1], la señal de voz ruidosa es realzada mediante un proceso iterativo de filtrados de Wiener, de respuesta frecuencial:

$$W(f) = \frac{Ps(f)}{Ps(f) + Pr(f)} \quad (1)$$

donde Pr(f) es el espectro del ruido estimado mediante un alisado del periodograma (método

Este trabajo ha sido financiado por la CICYT del MEC, bajo los proyectos TIC 95-1022-C05-02 y TIC 95-0884-C04-04, y por la CONSI+D de la DGA, bajo el proyecto P.64/96.

WOSA) en la zonas de silencio, y Ps(f) es el espectro de la señal de voz limpia, que no está disponible. Para solventar este problema se obtiene una estimación previa a partir de la señal ruidosa, procediéndose a una iterativa mejora de esta estimación a partir de la señal de voz realzada obtenida a la salida del filtrado precedente, como se muestra en la figura 1.



$$W_i(\omega)=\frac{P_{y_{i-1}}(\omega)}{P_{y_{i-1}}(\omega)+P_r(\omega)} \qquad \text{donde} \quad P_{y_{i-1}}(\omega)=\frac{g^2}{\left|1+\sum_{k=1}^{p}a_k e^{-j\omega k}\right|^2}$$

Figura 1. Esquema del algoritmo iterativo de Wiener clásico

En un primer momento, puede pensarse que se obtiene una mejora de la calidad de la señal realzada tras cada iteración, debido a que en cada una de ellas se hace uso de un modelado espectral AR de la señal de voz basada en una muestra de esta más limpia. Desafortunadamente aparecen otros factores negativos en el proceso que limitan las prestaciones y el número aconsejable de iteraciones del algoritmo. La señal filtrada presenta un ruido residual menor tras cada iteración pero también una mayor distorsión espectral. De este modo, un aumento del número de iteraciones no conlleva necesariamente un aumento de la calidad de la voz realzada. Es bien conocido que este algoritmo conduce a un estrechamiento ("picado") y un desplazamiento de los formantes de la señal de voz obtenida a medida que el número de iteraciones crece. Estos efectos pueden observarse en la figura 2. El efecto de picado espectral aparece aún cuando se utilice una estimación exacta del filtro de Wiener en cada iteración (es decir, aún cuando se suponga desacopladas la señal y el ruido a la salida de cada iteración), puesto que la solución MMSE siempre causa distorsión espectral de forma inherente. En [2] se muestra un detallado análisis de la convergencia de este algoritmo. Se prueba que el algoritmo converge hacia una estimación del filtro que tiende a cancelar todas las componentes de la señal que presentan una SNR inferior a 4.77 dB, e introduce una atenuación mayor (proporcional al nivel de ruido) que el filtro de Wiener óptimo en aquellas frecuencias con una SNR mayor que ese valor. Solo las frecuencias no contaminadas presentan atenuación nula como en el caso óptimo. Este análisis del comportamiento del algoritmo explica el picado del espectro que introduce.

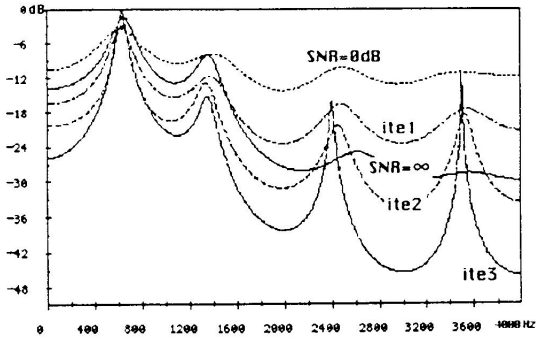


Figura 2. Evolución del efecto de "picado" de los formantes con el número de iteraciones. SNR=0 dB.

3. El algoritmo con criterio no cuadrático.

En la sección precedente se ha comentado la existencia de un compromiso entre el nivel de reducción de ruido y el nivel de distorsión espectral introducido por el algoritmo iterativo de Wiener clásico. El error de estimación obtenido en la salida del filtro de Wiener contiene dos componentes: un ruido residual y una componente de distorsión de señal. El segundo término, correlado con la voz, resulta mucho más perjudicial en la audición de la señal realzada que el segundo, incorrelado con la señal de voz. El criterio MMSE del algoritmo clásico de Wiener trata ambas componentes por igual. En el algoritmo propuesto se modifica el criterio o función de coste con objeto de adquirir control sobre este compromiso ruido residual - distorsión espectral de la señal. Así, se introducen dos términos que miden ambos componentes de error en la función de coste a minimizar. además, se introduce un término de esfuerzo o ganancia global del filtro con objeto de reducir los coeficientes del filtro con menor influencia en la minimización del resto de los términos. De este modo, la función de coste no cuadrática se define como:

$$L=\beta_1 E\left\{e^2(n)\right\}+\beta_2 E^2\left\{\sum_m y(n)r(n-m)\right\}+$$

$$+\beta_3 E^2\left\{\sum_m e(n)s(n-m)\right\}+\beta_4 E\left\{\sum_m w_m^2\right\} \qquad (2)$$

donde e(n) es el error residual, s(n) es la señal de voz limpia, r(n) es el ruido, y(n) es la señal de voz realzada y w_m son los coeficientes del filtro. Los factores β_i son de ponderación de los diferentes términos del a función de coste. El significado de los diferentes términos de la función de coste es:

- 1er término :error cuadrático (MSE) como en filtrado de Wiener clásico.
- 2º término: correlación cruzada entre señal realzada y ruido. Su minimización persigue

ortogonalizar la señal realzada y el ruido, y refuerza el objetivo de reducción de ruido del término MMSE.

-3º término: correlación cruzada entre error residual (distorsión de señal + ruido residual) y señal de voz. Su minimización persigue reducir la distorsión de señal y trabaja en sentido complementario al anterior.

-4º término: persigue la solución del filtro con coeficientes de norma L2 mínima consistente con la minimización de los otros tres términos. Tiende a eliminar los coeficientes que presentan una menor influencia en el proceso de minimización, reduciendo la varianza del error de estimación de estos, y en consecuencia, la magnitud de los lóbulos secundarios de la respuesta del filtro.

Se proponen dos tipos de soluciones al problema de minimización planteado: una, resuelta en el dominio de la frecuencia y equivalente a la solución propuesta en el algoritmo original de Lim-Oppenheim, correspondiente a un filtro ideal no causal; y dos, resuelta en el dominio del tiempo y que se corresponde con una solución causal directamente realizable. En ambos casos, el algoritmo iterativo se lleva a cabo como en el algoritmo original, de tal modo que la salida realzada de cada iteración se utiliza como señal limpia s(n) de la siguiente a efectos de estimación del nuevo filtro.

3.1 Solución no causal en el dominio de la frecuencia.

En este caso se trabaja en el dominio de la frecuencia, suponiendo un filtro idealmente no causal y de longitud infinita. La minimización directa de la función de coste L de la expresión (2) conduce a un filtro de respuesta frecuencial de expresión la siguiente:

$$W(f)=\frac{\beta_1 P_s(f)+\beta_3 P_s^2(f)}{\beta_1 [P_s(f)+P_r(f)]+\beta_2 P_r^2(f)+\beta_3 P_s^2(f)+\beta_4} \quad (3)$$

Esta expresión se reduce a la del Wiener clásico en el caso de escoger todos los parámetros de ponderación $\beta_i=0$ excepto $\beta_1=1$. En el caso general, el término $P_S^2(f)$ produce una sobreponderación de las frecuencias donde el nivel de señal es alto, proporcionando un valor de la respuesta frecuencial del filtro cercana a la unidad y previniendo así su distorsión. Por el contrario, el término $P_r^2(f)$ en el denominador produce una sobreponderación de las frecuencias con alto nivel de ruido, proporcionando un menor valor de la respuesta del filtro y, en consecuencia, una mayor cancelación de las mismas. La no realizabilidad del filtro es soslayada, como en el algoritmo clásico, mediante una aproximación basada en el muestreo frecuencial de la expresión exacta (2) haciendo uso de una FFT de orden N=256 y calculando los N coeficientes del filtro mediante FFT inversa. En realidad el filtrado se realiza directamente en el dominio de la frecuencia haciendo uso de una FFT de 512.

3.1 Solución causal en el dominio del tiempo.

La solución apuntada en la sección precedente usaba una sobreestimación de la longitud del filtro para prevenir el aliasing y el rizado inherente la método de diseño de filtros por muestreo en frecuencia. Un detallado estudio de la longitud real requerida por el filtro sugiere que un valor de N=21 sería suficiente. En el dominio del tiempo, la minimización de (2) conduce a unas ecuaciones normales generalizadas cuya solución responde a la expresión:

$$\underline{W}=[\beta_1 \underline{\underline{R}}_{xx}+\beta_2 \underline{\underline{R}}_{rr}^2+\beta_3 \underline{\underline{R}}_{ss}^2+\beta_4 \underline{\underline{I}}]^{-1}[\beta_1 \underline{P}_{ss}+\beta_3 \underline{\underline{R}}_{ss}\underline{P}_{ss}]$$
$$=\underline{\underline{R}}^{-1}\underline{P} \quad (4)$$

donde un subrayado doble significa matriz y uno simple significa vector. Los coeficientes de correlación de la señal de voz son estimadas desde la señal realzada obtenida en la iteración precedente y los del ruido a partir de las zonas de silencio. Esta solución requiere una alta carga computacional, y además, la inversión de una matriz de tal dimensión puede producir problemas numéricos. Por ello, se prefiere el uso del algoritmo de Steepest Descent (SD), donde el mínimo es obtenido por un algoritmo de gradiente iterativo de expresión:

$$\underline{\nabla}_w(n)=\underline{\underline{R}}\,\underline{W}(n)-\underline{P}.$$

4. Resultados preliminares.

En esta sección se presentan algunos resultados comparativos obtenidos haciendo uso del último algoritmo propuesto en le dominio del tiempo y del algoritmo SD de gradiente. En la figura 3 se muestra la envolvente espectral LPC de una trama de voz realzada haciendo uso de a) algoritmo clásico de Lim con hasta tres iteraciones, y b)algoritmo propuesto con dos iteraciones y diversas combinaciones de parámetros de ponderación β_i. La relación señal-a-ruido (SNR) global es de 9 dB. Como puede apreciarse, El efecto de picado de los formantes, de intensidad creciente con el número de iteraciones, es evidente en el caso el algoritmo clásico de Lim. Por el contrario, este efecto de

picado es muy bajo en el algoritmo propuesto, especialmente en el caso de $\beta_1=\beta_3=0,5$, y el seguimiento de la envolvente espectral es muy alto, incluso en los valles. Los resultados con el algoritmo SD son mejores aún en el caso de criterio MMSE puro. Los test de audición realizados confirman estos resultados, obteniéndose una calidad superior con los algoritmos propuestos. Con respecto a medidas objetivas, no siempre correladas con los test subjetivos, se obtiene una mayor mejora de la SNR en el caso del algoritmo clásico, al menos tras la primera iteración (tras diversas iteracicnes, el efecto de picado espectral llega a ser dominante y la SNR decrece rápidamente) Por el contrario, en términos de distancia espectral se obtiene mejores resultados con el algoritmo propuesto, y en concordancia con los resultados mostrados en la figura 3. Resultados adicionales serán aportados en la presentación del trabajo.

5. Conclusiones

Se ha propuesto un nuevo algoritmo de realce de señal de voz basado en un filtrado iterativo. El diseño del filtro se hace en base a la minimización de una función de coste no cuadrática, que incluye, además del clásico término MSE, dos términos de correlación cruzada error-señal. Además se incluye un término de norma cuadrática L2 de los coeficientes del filtro que representa la ganancia global o esfuerzo del mismo. Los dos términos de correlación cruzada controlan el valor de los dos componentes del error de estimación: ruido residual y distorsión de la señal. Con el término de norma L2 de los coeficientes del filtro se intentan reducir el valor de aquellos coeficientes que no tienen un papel representativo en la minimización del error de estimación, reduciendo la varianza del error de estimación de estos, y en consecuencia, la magnitud de los lóbulos secundarios de la respuesta del filtro. Para resolver el problema de minimización planteado se proponen dos tipos de soluciones: una, en el dominio de la frecuencia, proporciona la respuesta frecuencial ideal correspondiente a un filtro no causal de longitud infinita; la otra, se resuelve en el dominio del tiempo y proporciona la respuesta impulsional de un filtro realizable, causal y de longitud finita. En este último caso se proporcionan dos variantes diferentes: una, que conduce a la inversión de una matriz de dimensión alta como solución a un sistema generalizado de ecuaciones normales, y otra, que resuelve este sistema mediante un algoritmo iterativo de tipo gradiente o Steepest Descent (SD). Los algoritmos propuestos trabajan bien y los resultados obtenidos superan los correspondientes al caso de algoritmo iterativo clásico MSE (Wiener), tanto en lo que respecta a distancias espectrales como a calidad subjetiva de audición. Se ha observado que la mejor elección corresponde al algoritmo en el dominio del tiempo basado en el SD.

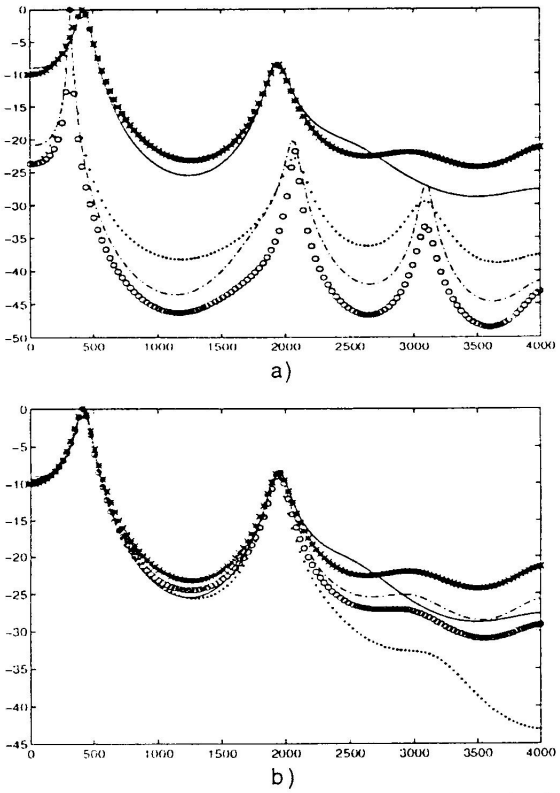


Figure 4. Envolventes espectrales de la voz realzada. SNR de la voz ruidosa= 9 dB.. a) algoritmo clásico Línea continua:voz limpia.'*':voz ruidosa. '.','o' y '-.': voz realzada tras 1,2 y 3 iteraciones, respectivamente. b) Algoritmo propuesto. Línea continua: voz limpia. '*': voz ruidosa. '.':β_1=1, β_3=0; 'o':β_1=0.7, β_3=0.3; '-.':β_1=0.5, β_3=0.5. En todos los casos $\beta_2=\beta_4=0$.

Referencias

[1] J.S.Lim and A.V.Oppenheim, "All-Pole Modeling of Degraded Speech". IEEE Trans ASSP, pp197-210.June 1978.

[2] E.Masgrau et al ,"Speech Enhancement by Adaptive Wiener Filtering based on Cumulant AR Modelling". ESCA Workshop on Speech Processing in Adverse Conditions, pp 143-146. Cannes, France. November 1992.

[3] J.Salavedra,E.Masgrau.et al. "A Speech Enhancement System using Higher-order AR estimation in real environments". EUROSPEECH'93, pp. 223-226. Berlin, Germany. September 1993.

[4]J.Salavedra,E.Masgrau,et al."Some Robust Speech Enhancement Techniques using Higher-Order AR Estimation". EUSIPCO-94. Edinburgh, U.K. September 1994.