Codificación APVQ-extendida de Voz de Banda Ancha

Enrique Masgrau*, Josep Salavedra**
* Departamento de Ingeniería Eléctrica e Informática. CPS Ingenieros.
Universidad de Zaragoza. C/María de Luna,3. 50015-ZARAGOZA.

Tfno: 976 519892, Fax: 976 512932, E-mail: masgrau@mcps.unizar.es
** Departamento de Teoría de la Señal y Comunicaciones. UPC.
ETSI Telecomunicación. Aptdo. 30.002. 50071-BARCELONA
Tfno: 93 4016754, Fax: 93 4016447, E-mail: mia@tsc.upc.es

Abstract.- This paper describes a coding scheme for broadband speech. It can be seen as a vectorial extension of an conventional ADPCM encoder. In this scheme, the vector signal is formed with one sample of the normalizaed prediction error of each subband and then, it is vector quantized. It combines the advantages of the scalar prediction and of the vector quantization (VQ). We handle the high vector dimensionality by using a multi-VQ. It requires a previous subvector division and an adequate bit assignement among them. This scheme shows an high capacity to drive large dynamic range signals like broadband speech.

1.- Introducción

La combinación de las técnicas de división en subbandas con cuantificación vectorial y predicción adaptativa proporciona muy buenos resultados en codificación de señal de voz de banda estrecha (4 KHz.) a velocidades medias de 1 bit/muestra (8 Kbps.). Un ejemplo de este tipo de codificadores es el denominado APVQ (Adaptive Predictive Vector Quantization) [1] que consiste, básicamente, en una división de la señal de voz en 8 subbandas de 500 Hz cada una mediante un banco de filtros QMF, seguido de una cuantificacion ADPCM "backward" de cada una de las bandas, con la particularidad de que la cuantificación del error de predicción en cada una de las subbandas se realiza mediante una cuantificación vectorial (VQ), de tal modo que cada uno de estos errores de predicción constituye una de las componentes del vector de entrada al VQ. Es decir, en vez de cuantificar el error de predicción de cada una de las subbandas de forma independiente mediante un cuantificador escalar, se cuantifican en bloque mediante un VQ. Además, este VQ es adaptativo, en el sentido de que los errores de predicción son previamente normalizados en ganancia mediante una estimación "backward" de ésta, o lo que es lo mismo, se hace uso de una cuantificación VQ ganancia-forma adaptativo.

Como se detalla en la referencia [1], a esta velocidad de transmisión moderada de 1 bit/muestra, la predicción adaptativa no aporta ninguna ventaja en las bandas 5 a 8 (por encima de 2 KHz.), con lo que puede prescindirse de ella en estas bandas. Ello es debido a que el error de cuantificación producido en la representación de la señal en cada una de las subbandas enmascara el potencial de blanqueado en tiempo proporcinado por la predicción, ya bastante reducido debido a la división en subbandas, que como es bien sabido proporciona un blanqueado en frecuencia. En las figuras 1 y 2 se muestran los esquemas generales del codificador y del decodificador APVQ, respectivamente.

En este trabajo se presenta la extensión de este codificador al caso de voz de ancho de banda de 7 KHz., es decir, de calidad conversacional, adecuada para aplicaciones multimedia. En este caso, los requerimientos de calidad obligan a trabajar a velocidades de 1,5 a 2 bits/muestra (de 24 a 32 Kbps con una frecuencia de muestreo de 16 Khz.).

En este caso, el número de subbandas en que se divide el margen de 0 a 8 Khz de la señal es de 16, siendo todas ellas de 500 Hz. de anchura, y despreciandose las dos subbandas superiores debido a su despreciable contenido energético. En este caso, la mayor precisión de representación de las muestras en cada subbanda proporciona un mejor aprovechamiento del potencial de blanqueado de la predicción adaptativa, lo que aconseja el uso de ésta por lo menos en las 8 primeras subbandas, las de mayor contenido energético.

Por otro lado, la cuantificación vectorial de las 14 componentes correspondientes a cada una de las subbandas útiles no puede realizarse en bloque, ya que el costo computacional es inabordable: ¡cuantificacion

Este trabajo ha sido financiado por el Plan Integrado de Banda Ancha (PLANBA) perteneciente al Plan Nacional de I+D.

VQ de un vector de dimensión 14 mediante un codebook de 2^{21} a 2^{28} palabras código (correspondientes a velocidades entre 1,5 y 2 bit/muestra)!. Para soslayar este problema se hace uso de un cuantificador multi-VQ, consistente es la división del vector total en varios subvectores de dimensiones adecuadas para que su cuantificación requiera una complejidad moderada. Estos subvectores, y su consiguiente cuantificación, puede, definirse de dos formas diferentes: 1) dimensiones fijas de los subvectores, con una asignación dinámica de bits entre éstos; 2) dimensiones variables de los subvectores tal que sus energías sea lo mas uniformes posible y una asignación uniforme de bits. En ambos casos, la asignación de bits es basada en las estimaciones backward de la energía o ganancia de cada subbanda, disponible en el codificador y en el decodificador, lo que no requiere uso de información lateral. Además, una mejora de la calidad subjetiva de la voz puede obtenerse mediante un conformado espectral de ruido, el cual es obtenido introduciendo una ponderación en la ganancia de cada una de las subbandas, lo que equivale a una ponderación espectral.

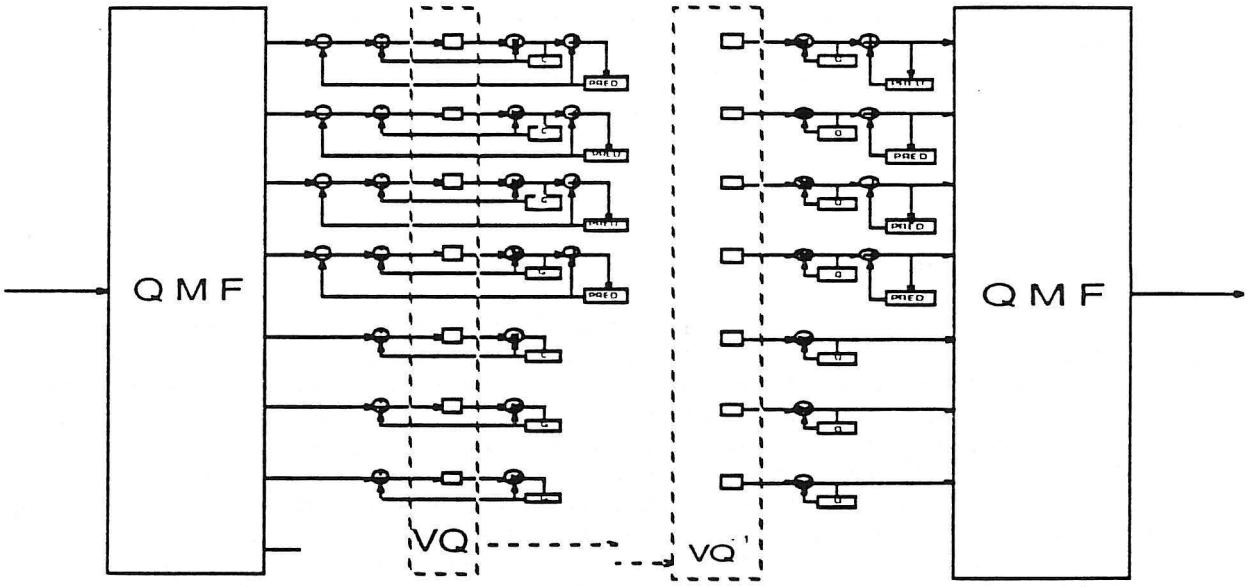

Figura 1. Esquemas generales de un codificador y decodificador APVQ. Caso de división en 8 subbandas: las cuatro bandas inferiores incluyen predicción adaptativa, las 3 siguientes la excluyen y la banda superior no es transmitida.

2.- Predicción adaptativa

Los predictores utilizados en la codificación de cada una de las subbandas son de tipo FIR, adaptativos y de tipo "backward" tanto porque el algoritmo de adaptación se basa en la muestras reconstruidas como porque la predicción misma se basa en esas mismas muestras. Por ello, no es necesario transmitir información lateral sobre el cálculo o valores de los coeficientes de los predictores. Los algoritmos de adaptación connsiderados son el conocido LMS y el GAL (Gradiente Adaptive Lattice) [1,4]. El segundo, más complejo que el LMS y, por consiguiente, de mayor costo computacional, proporciona unas prestaciones mucho mejores debido a su mayor velocidad de convergencia, sobre todo trabajando con alta calidad de codificación [1], como es el caso que nos ocupa. Por ello, este algoritmo GAL es el elegido, pues la mejora de calidad proporcionada compensa el mayor costo computacional requerido. Respecto al orden de los predictores a utilizar, cabe distinguir entre las bandas inferiores, donde la estructura de tipo periódico (rayas espectrales) del espectro en los intervalos sonoros de la voz es nítida, y las superiores, donde esta estructura se pierde en gran medida. En el primer caso, es adecuado utilizar longitudes mayores, de tal modo que se pueda aprovechar la periodicidad de la señal en los intervalos sonoros. Esta longitud no tiene por qué ser demasiado alta, debido a que el diezmado de las subbandas reduce la longitud del periodo de pitch por el mismo factor de diezmado. Longitudes de valor 9 ó 10 son las que proporciona un buen equilibrio entre prestaciones y complejidad (no debe olvidarse que las prestaciones de los algoritmos adpatativos, y por consiguiente la capacidad de predicción de los predictores que aquéllos gobiernan, se deterioran con el número de sus

coeficientes o longitud). Para los predictores correspondientes a las bandas superiores, el orden se reduce drásticamente, llegando a eliminarse la existencia misma del predictor en las bandas más altas.

3.- Normalización adaptativa de la ganancia

El error de predicción de cada subbanda (o la señal misma, si no se usa predicción como es el caso de las bandas más altas), se normaliza previamente a su cuantificación, tal y como se muestra en la figura 1. Esta normalización permite reducir el margen dinámico de la señal a cuantificar, lo que mejora la calidad de ésta y proporciona robustez frente a cambios de nivel en la potencia de la señal a codificar. Es adecuado destacar que esta normalización del nivel de señal se efectúa de forma independiente en cada subbanda o componente del vector a cuantificar, lo que permite adecuar el cuantificador VQ a los diferencias relativas de potencia entre subbandas. Es decir, el vector presentado al VQ es un vector normalizado por un factor que tendrá que ser tenido en cuenta en el diseño del codebook del VQ, ya que el error de cuantificación producido en cada componente se vera luego magnificado (o reducido) por este factor de ganancia. Por otro lado, este factor de ganancia no es necesario que sea transmitido al receptor, pues, como en el caso de los coeficientes del predictor, su cálculo es realizado en modo "backward" basado en señales disponibles en el receptor, y por consiguiente, reproducible en éste. El algoritmo de estimación del factor de ganancia de cada subbanda puede realizarse mediante una simple estimación recursiva de un sólo polo (alisado por una ventana exponencial) o mediante una más sofisticada técnica, que hace uso de un predictor adaptativo "backward" que predice o estima la ganancia de la componente actual basándose en las ganancias de las componentes codificadas anteriores. El uso de este segundo método está más indicado cuando el nivel de calidad de la codificación es alto, como es el caso de codificación de banda ancha que nos ocupa. De todos modos, la simplicidad y buenas prestaciones ofrecidas por el metodo de estimación recursiva suele hacerlo más atractivo en la realización práctica de estos sistemas de codificación, ya de por sí bastante complejos.

4.- Cuantificador vectorial multi-VQ.

Como ya se ha comentado en la introducción, la alta dimensionalidad del vector a cuantificar y el altísimo número de palabras código que contendrían los codebooks de un único VQ, requieren del uso de un multi-VQ. Esto consiste en segmentar el vector de dimensión k=14 en varios subvectores de dimensiones k_i más reducidas. De este modo, la cuantificación de cada subvector se realiza de forma independiente, de tal modo que el VQ se comporta como un multi-VQ o código producto. Obviamente, esta solución es subóptima, pero la degradación sufrida no es demasiado importante si se eligen cuidadosamente la segmentación y la asignación de bits a cada uno de ello. De hecho, se han obtenido muy buenos resultados con el uso de estás ideas en codificación de voz de banda estrecha usando codificadores vectoriales transformados [2,3]. En definitiva, se trata de introducir un compromiso prestaciones-complejidad, que resulta, sin duda, muy provechoso. El resto del sistema APVQ permanece básicamente invariante respecto al esquema general, siendo necesario únicamente una adaptación de los parámetros de diseño a la nueva situación. Si se pretende no aumentar demasiado la complejidad del VQ, parece claro que las subbandas inferiores, a las que habrá que asignar un número promedio elevado de bits/subbanda, deben agruparse en subvectores de dimensiones más cortas. Por contra, las bandas superiores permitirán la definición de subvectores de dimensiones más altas.

El objetivo está bien definido: encontrar la mejor segmentación y asignación de bits posible que proporcionen el mejor comportamiento sin trasgredir el máximo de complejidad en ninguno de los cuantificadores VQ que componen el código producto multi-VQ. Es decir, se define un valor $C_{máx}$ del producto $k_i\, 2^{k_i r_i}$ (complejidad del VQ), donde r_i es el número medio de bits/muestra asignado al subvector i, y para esta cota se buscan los pares (k_i, r_i) que proporcionen la mejor SNR. Valores típicos de $C_{máx}$ sueden tomarse en el orden de 1000 a 3000. Para la consecución del objetivo marcado pueden seguirse dos estrategias: 1) obtener la mejor segmentación k_i en subvectores, usando una extensa base de datos de estrenamiento, y luego realizar una asignación dinámica de bits a estos subvectores con criterios de optimidad; y 2) definir un número de subvectores, es decir su componente inicial y final (o su inicio y su dimensión k_i), tal que la energía de todos los subvectores sea lo más homogénea posible, y entonces asignar el mismo número de bits a cada uno de ellos. La primera resulta más sencilla en lo que se refiere al entrenamiento y almacenamiento de los diferentes VQ correspondientes a cada subvector, pero más complicado el cálculo de la asignación dinámica de bits. La segunda, por contra, resulta más compleja en lo que se refiere a definición de los

subvectores (requieren dos índices: muestra inicial y dimensión) y el entrenamiento de los respectivos VQ, pero resulta muy sencilla en lo que se refiere a la asignación de bits, que es uniforme e invariante.

En la primera estrategia, se parte de una asignación dinámica de bits óptima basada en resultados de la teoría de distorsión-transmisión:

$$r_i = r + \beta_i + \frac{1}{2}\log_2 \frac{\left(\prod_{j=1}^{k_i} \sigma_{ij}^2\right)^{1/k_i}}{\left(\prod_{j=1}^{m}\prod_{h=1}^{k_i} \sigma_{jh}^2\right)^{1/k}}$$

donde r es el número medio de bits/muestra, r_i es el número medio de bits/muestra asignado al subvector i, k_i es la dimensión del subvector i, m es el número de subvectores y σ_{ij}^2 es la energía promedio de la componente de señal j correspondiente al subvector i. El término β_i modifica la asignación de bits al tener en cuenta las ventajas de la dimensionalidad en la cuantificación vectorial y depende de k_i. En el caso escalar toman todos el valor nulo; como resulta difícil de calcular, se suele proceder como si fuera el caso escalar (β_i=0). Se procede como sigue: se trabaja con una base de datos suficientemente amplia y sobre ella se asigna a cada componente el número de bits que corresponda para el caso escalar (m=k, k_i=1), según la fórmula anterior. a continuación, se define el valor k_1 que hace que el par (k_1,r_1) proporcione el valor $C_{máx}$ (menor o igual), y se procede del mismo modo con el resto de subvectores. Con los subvectores de la base de datos asignados a cada clase se diseñan los VQ de diferentes tamaños, que serán utilizados en el proceso de codificación propiamente dicha. Los bits sobrantes debido a la truncación se asignan a los VQ en los que producen una mayor mejora, generalmente correspondientes a los subvectores de mayor energía promedio. En el modo de codificación se hace uso de las dimensiones de los subvectores obtenidas en el proceso de diseño, en base a ellas y a las energías estimadas para cada subvectors se calcula la asignación de bits correspondiente a cada uno de ellos y se hace uso del VQ del tamaño correspondiente. No se requiere información lateral, ya que el cálculo de la asignación de bits es "backward".

En la segunda estrategia, se procede sobre la base de datos de entrenamiento, de modo que se van definiendo las dimensiones de los m subvectores distintos, comenzando con m=1, de tal forma que la energía de cada uno de ellos sea aproximadamente del orden de σ^2/m, donde σ^2 es la energía media total del conjunto de todas las k subbandas. La máxima complejidad $k_i\, 2^{(r.k/m)}$ obtenida debe ser inferior a $C_{máx}$. En caso de superarse esta cota se aumenta el número de subvectores m en una unidad y se actúa del mismo modo. Como ya se comentó, la no invarianza del índice de inicio de los diferentes subvectores complica bastante el entrenamiento de los diferentes VQ. En el modo de codificación se procede de la misma forma, pero con el número de subvectores m ya fijado por el proceso anterior de diseño.

Es conveniente introducir una ponderación espectral del ruido de cuantificación a fin de mejorar la calidad subjetiva de la voz. Esta ponderación espectral, que busca mantener siempre el nivel de ruido por debajo del de la señal en todas las frecuencias, se consigue reduciendo el peso de las energías σ_{ij}^2 (o mejor, de su estimación g_i^2) en la asignación de bits a los correspondientes subvectores. Esto se consigue sustituyendo estas energías σ_{ij}^2 por $\sigma_{ij}^{2\gamma}$ a efectos del cálculo de la asignación de bits. Valores de γ del orden de 0,7 a 0,8 proporcionan resultados adecuados. Más detalles sobre las características de este tipo de ponderación se dan a continuación al hablar de la ponderación espectral dentro de cada subvector.

Una vez resuelta la asignación de bits a cada uno de los subvectores del multi-VQ, veamos la cuantificación vectorial (VQ) decada uno de éstos. Esta cuantificación presenta las dos características especiales de ser adaptativa, a través de la normalización dinámica de cada una de los componentes por una estimación adaptativa de su respectiva ganancia anteriormente comentada, y de hacer uso de una distancia dinámica espectralmente ponderada, que proporciona un conformado espectral del ruido de cuantificación, dentro de cada subvector, y la correspondiente mejora de la calidad subjetiva de la señal codificada. Este conformado de ruido, denominado intra-subvectorees, se añade al denominado conformado de ruido inter-subvectores, que ya ha sido comentado anterioemente al hablar de las políticas de asignación dinámica de bits entre subvectores. La ponderación espectral intra-subvectorse se obtiene utilizando en el proceso de cuantificación (y también en el de diseño del VQ) de una medida del tipo:

$$D = \sum_i w_i(n)\,(e_i(n) - e_{qi}(n))^2 = \sum_i w_i(n)\, q_i^2(n)$$

donde $w_i(n)$ es el peso del iésimo componente del vector (o error de predicción de la subbanda i), y $e_i(n)$ y $e_{qi}(n)$ son el error de predicción y éste cuantificado, respectivamente. La diferencia entre estas dos últimas magnitudes, $q_i(n)$, es el error de cuantificación final obetnido como producto de la codificación de la señal en cada subbanda. Si $w_i(n)$ fuera contante con i (p.e., unidad), el error de cuantificación tendería a ser blanco, es decir, la varianza de éste, $\sigma_{qi}^2(n)$ sería cte con i. Con $w_i(n)$ no constante, lo que resulta ser blanco es el producto $w_i(n).\sigma_{qi}^2(n)$ = cte. Una elección adecuada es la de tomar el factor de ponderación de forma similar al caso de ponderación inter-subvectores:

$$w_i(n) = w_{0i}.(g_i^2(n))^{\lambda} \qquad -1 < \lambda < 0$$

donde w_{0i} es un término fijo de ponderación, diferente para cada subbanda, y $g_i^2(n)$ es la estimación "backward" de la varianza o ganancia del error de predicción, disponible en el cuantificador. Puede comprobarse [4] que esta ponderación proporciona una SNR_i de valor:

$$SNR_i = K + 10.\log(w_{0i}) + \lambda.10.\log(g_i^2) + 10.\log(\sigma_i^2)$$

donde K es una contante que depende de la potencia y del espectro de la señal y del valor de λ. Es decir, para λ=0, el ruido de cuantificación tiende a ser plano (óptimo) y se obtiene una SNR en cada subbanda que es proporcional a la potencia de la señal en dicha banda y a la ponderación fija escogida en ella. Asimismo, para valores negativos de λ, la SNR tiende a ser similar para todas las bandas, el ruido de cuantificación tiene una envolvente espectral semejante a la de la señal y la SNR_i de cada banda se deteriora tanto menos cuanto menor sea su energía, medida por el valor de g_i^2. Ello proporciona una forma de redistribuir entre bandas el nivel de ruido de cuantificación, objetivo final del conformado de ruido. Valores cercanos a λ=-0.3 resultan ser los más adecuados. Como ya se comentó anteriormente, esta ponderación espectral debe ser tenida en cuenta en el diseño de los centroides o palabras código de cada uno de los VQ.

5.- Conclusiones

El codificador aquí presentado, denominado APVQ-extendido, resulta adecuado para la codificación de calidad de señal de voz de banda ancha de 7 KHz., cuya característica principal es el alto margen dinámico espectral que presentan este tipo de señales. Este sistema, al incluir división en subbandas y cuantificación vectorial adaptativa, presenta muy buenas propiedades para manejar con altas prestaciones este tipo de señales. Los resultados previos obtenidos, junto a las buenas prestaciones obtenidas con este sistema para señales de voz de banda estrecha (4 Khz.) y a las propiedades intriínsecas comentadas del sistema, garantizan un muy buen comportamiento con señales de voz de banda ancha. Una evaluación empírica del sistema será presentado en la lectura de este trabajo.

Referencias

[1] Enrique Masgrau, , J.B. Mariño. "Subband splitting, adaptive scalar prediction and vector quantization". . Proc. EUSIPCO 88. Grenoble

[2] Enrique Masgrau, J.A.R. Fonollosa, J.R. Mallafre. "Predictive SVD-transform coding of speech with adaptive vector quantization". Proc. IEEE ICASSP 91. Toronto.

[3] T. Moriya, M. Honda. "Transform coding of speech with weighted vector quantization". Proc. IEEE ICASSP 87. Dallas.

[4] J.A.R. Fonollosa. "Cuantificación vectorial adaptativa aplicada a la codificación de voz". Tesis Doctoral. Univ. Polit. de Cataluña UPC. Julio.