Comparación entre la Aplicación de la DCT y la KLT a la Compresión de Audio (Enero 2012)

Iván López Espejo

Documento donde se presenta la metodología y resultados derivados de la aplicación de la DCT y la KLT a la compresión de la señal de audio.

I. Introducción

Tanto a la hora de almacenar como de transmitir una señal, resulta trascendental una correcta codificación de la misma con el fin de llevar a cabo una operación eficiente. En este sentido, muy importante resulta la compresión de los datos, de tal forma que se encuentre un compromiso entre las dos características deseables que son la no pérdida de información relevante y un pequeño volumen con el fin de que se minimicen los recursos dedicados a su almacenaje o transmisión.

Por ello, en el presente trabajo abordamos la comparación entre la aplicación de la DCT y la KLT a la compresión de audio. Estudiamos las características de sendas transformaciones y comparamos los resultados relativos en términos de compresión que ofrecen tras la puesta a cero del conjunto de coeficientes de menor magnitud en el dominio transformado para un fragmento de audio de test.

II. Introducción a la compresión con DCT

Nuestro propósito a la hora de emplear una transformada ortogonal para representar un fragmento sonoro en un nuevo dominio es el de estar en condiciones de lograr un alto grado de compresión de datos sin sacrificar apenas calidad auditiva.

La DCT nos permite calcular las componentes de proyección de la señal en el dominio original sobre el espacio de transformación compuesto por las funciones base tipo coseno. En otras palabras, la señal se puede expresar como combinación lineal de un conjunto de funciones que definen una base en el espacio de transformación. Debido a que la DCT es una transformada ortogonal, los coeficientes de transformación (las funciones base tipo coseno) cumplen con las propiedades necesarias de ortonormalidad y completitud. La propiedad de completitud asegura que la transformada es invertible y la de ortonormalidad garantiza que la energía se localiza en los primeros coeficientes de la transformación, lo que equivale a decorrelar las componentes del vector de señal de entrada en mayor o menor grado, cosa que es mucho más eficiente en el dominio transformado.

La ecuación de análisis o transformada directa se define como:

$$X(k) = \alpha_k \sum_{n=0}^{N-1} x(n) \cos\left(\frac{\pi k(2n+1)}{2N}\right),$$

donde:

$$\alpha_k = \begin{cases} \frac{1}{\sqrt{N}} & si\ k = 0 \\ \sqrt{\frac{2}{N}} & si\ k \neq 0 \end{cases}.$$

Por otro lado, la ecuación de síntesis o transformada inversa se define de la forma:

$$x(n) = \alpha_k \sum_{k=0}^{N-1} X(k) \cos\left(\frac{\pi k(2n+1)}{2N}\right).$$

Como ya hemos expuesto, la DCT concentra mucho la información en los primeros coeficientes, por lo que podemos aplicar técnicas para reducir la cantidad de datos en el dominio transformado sin perder apenas información relevante que nos lleve a una pérdida notable de la calidad del fragmento sonoro en el dominio temporal.

En el presente trabajo, para llevar a cabo la compresión, se aplica la técnica de retención por umbralización sobre las sucesivas tramas transformadas, que consiste en mantener únicamente aquellos coeficientes en el dominio transformado que superen, en magnitud, un cierto umbral, poniendo el resto de ellos a cero. No obstante, para llevar a cabo una comparativa en función del factor de compresión, se opta por poner a cero, por cada trama transformada, los n coeficientes de menor magnitud con el fin de conseguir un factor de compresión del $(n/N) \times 100\%$, donde $n < N$ y N es el tamaño de la trama sobre la que se aplica la transformación. Una técnica similar de cuantización es la empleada en la codificación de imágenes JPEG, cuyo diagrama de bloques se puede observar en la figura 1. Notar cómo la cuantización se lleva a cabo en el dominio transformado tras aplicar la DCT, sucesivamente, sobre bloques de imagen de 8x8 píxels.

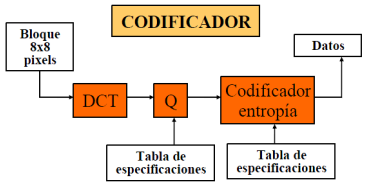

Fig. 1. Diagrama de bloques de un codificador JPEG.

III. Formulación de la transformada de Karhunen-Loève

Como hemos esbozado, la transformada óptima será aquella que decorrela completamente los píxels en el dominio transformado. Esta transformada recibe el nombre de transformada de Karhunen-Loève, que no es más que la herramienta estadística del análisis de las componentes principales (PCA).

Sea $\boldsymbol{x}$ un vector compuesto por N variables aleatorias y sea $\boldsymbol{y} = A\boldsymbol{x}$ el vector en el dominio transformado. Se trata, por tanto, de encontrar la matriz A que permita que $\boldsymbol{y}$ se encuentre totalmente decorrelado. Esto implica que la matriz de covarianza de $\boldsymbol{y}$ es diagonal. Si suponemos que trabajamos con procesos aleatorios de media nula (como típicamente es el caso de la señal de audio), la matriz de covarianza de $\boldsymbol{y}$ se puede expresar como sigue:

$$\Sigma_y = E[\boldsymbol{y}\boldsymbol{y}^T] = E[A\boldsymbol{x}\boldsymbol{x}^T A^T] = AE[\boldsymbol{x}\boldsymbol{x}^T]A^T = A\Sigma_x A^T,$$

donde Σ_x es la matriz de covarianza del proceso aleatorio descrito por $\boldsymbol{x}$. Como hemos dicho, deseamos que el proceso aleatorio en el dominio transformado se encuentre máximamente decorrelado, por lo que Σ_y es una matriz diagonal, de tal forma que, por el teorema de descomposición espectral de matrices simétricas, ya que sabemos que Σ_x es una matriz de coeficientes reales y simétrica (matriz de covarianza), A es una matriz ortogonal compuesta por los autovectores por filas asociados a los autovalores de la matriz de covarianza del proceso aleatorio en el dominio original. Por tanto, dichos autovectores conforman una base ortogonal, la cual define la base del espacio de transformación óptimo de Karhunen-Loève. Por todo esto, $L = [\boldsymbol{l}_1, \boldsymbol{l}_2, \dots, \boldsymbol{l}_N]$ es la matriz de transformación óptima de KL compuesta por los autovectores de Σ_x dispuestos por columnas, la cual se relaciona con A de la forma $A = L^T$.

Para garantizar mínima distorsión es importante que los autovalores en la matriz diagonal estén ordenados descendentemente $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N)$, hecho que condiciona la posición de los autovectores en la matriz de transformación. Notar cómo los autovalores no son más que las varianzas de las variables aleatorias del proceso original. Por todo esto, podemos expresar:

$$\Sigma_y = L^T E[\boldsymbol{x}\boldsymbol{x}^T]L = L^T \Sigma_x L = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}.$$

Es importante notar cómo las funciones base para la transformación no están definidas a priori sino que dependen de la señal en concreto que se vaya a transformar, lo que ya manifiesta una desventaja respecto de la DCT, y es que no existe un algoritmo de transformada rápida como sí ocurre para esta última. Principalmente por este hecho, el coste computacional es más severo para KLT en comparación con la DCT.

Para el caso de la señal de audio, podemos y se ha optado por definir N variables aleatorias diferentes donde N es el tamaño de trama de transformación, de tal forma que las sucesivas tramas conforman realizaciones de las N variables aleatorias. De esta forma, si el fragmento sonoro se compone de K muestras y K es múltiplo del tamaño de trama N, tendremos N variables aleatorias y K/N realizaciones. Sea la matriz simétrica de covarianza de dichas variables aleatorias $\Sigma_{\mathrm{x}} = \mathrm{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}]$, sus autovalores se calculan a partir de encontrar las raíces del polinomio característico, $p(\lambda) = |\Sigma_{\mathrm{x}} - \lambda \mathrm{I}| = 0$, y los autovectores que definirán la matriz de transformación L se calculan como $\Sigma_{\mathrm{x}} \boldsymbol{l}_n = \lambda_{\mathrm{n}} \boldsymbol{l}_n \ \forall n \in [1, N]$, donde $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_{\mathrm{N}}]$. Una vez hecho, por cada trama sonora se puede aplicar la transformación como:

$$Y = L^T X,$$

donde Y equivaldría a la trama sonora en el dominio transformado. Tras realizar las operaciones deseadas de compresión, podemos aplicar la transformada inversa sobre Y para volver al dominio original como:

$$X = LY.$$

Esto es posible gracias a que, como dijimos, las matrices de transformación compuestas por los autovectores son ortogonales y, por tanto, se verifica que $L^{-1} = L^T$. Finalmente comentar que, como comprobaremos en la práctica, el rendimiento de ambas transformadas es similar, por lo que puede ser más interesante a efectos computacionales emplear la DCT como realmente los estándares prefieren sobre KLT.

IV. Desarrollo

Para el testeo del rendimiento relativo de la DCT y la KLT en el ámbito de la compresión a través de la técnica de retención por umbralización se ha seleccionado un fragmento de test en formato *.wav* originalmente estéreo muestreado en calidad CD-Audio (44100Hz). Dicho fragmento sonoro se preprocesó para trabajar con un solo canal (mono), de la forma:

$$x(n) = \frac{x_L(n) + x_R(n)}{2},$$

donde $x_L(n)$ representa el canal izquierdo del par estéreo y $x_R(n)$ el derecho. El nuevo vector de muestras de audio se remuestrea a 48000Hz con el fin de poder utilizar posteriormente una implementación de PEAQ (Perceptual Evaluation of Audio Quality), un método para la medida objetiva de la calidad del audio en términos perceptuales descrito en el estándar ITU-R BS.1387. El resultado de esta operación es el fragmento sonoro de referencia con el que finalmente se llevan a cabo las pruebas, por lo que es exportado a un fichero *.wav*. Además, debido a que $x(n)$ (una vez remuestreado) tiene una longitud de $K = 592771$ muestras, por sencillez para aplicar las transformaciones se escoge un tamaño de trama divisor entero de K, que es para la ocasión $N = 613$ muestras. De este modo, tanto la

transformada discreta del coseno como la transformada de Karhunen-Loève se aplican un total de $K/N = 967$ veces, eliminando en cada caso los n coeficientes transformados de menor magnitud, donde $n < N$ y se ajusta para conseguir un factor de compresión del $(n/N) \times 100\%$.

Tras calcular la matriz de transformación óptima para la KLT en los términos explicados en la introducción teórica, se aplican ambas transformadas a cada una de las tramas y, para cada una de ellas, se hacen cero los n coeficientes de menor magnitud en el dominio transformado.

A continuación se muestra un pequeño ejemplo gráfico de funcionamiento para el caso de compresión con un factor del 90% (es decir, por cada trama hacemos cero en el dominio transformado los 552/613 coeficientes de menor magnitud). La figura 2 muestra el segmento original que se va a comprimir en el dominio temporal.

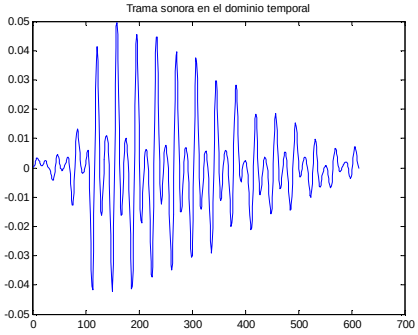

Fig. 2. Trama sonora en el dominio temporal que será comprimida.

De otro lado, las figuras 3 y 4 muestran los coeficientes transformados para el caso de la DCT y la KLT, respectivamente, de la trama anterior. Lo primero que llama la atención, tal y como se dijo, es que ambas (aunque sólo lo hace máximamente la KLT) decorrelan en mayor o menor grado los datos en el dominio transformado, pudiéndose observar que los coeficientes de mayor magnitud se agrupan en las primeras posiciones. Sobre el conjunto de coeficientes transformados (en color azul) se superpone el resultado de establecer a cero los 552 coeficientes de menor magnitud (en rojo). Sobre el resultado de esta manipulación se aplica la transformada inversa para obtener la trama sonora en el dominio original comprimida.

Por último, en la figura 5 se puede apreciar una ampliación cualquiera sobre un conjunto de muestras en el dominio temporal donde se comparan la señal original, la comprimida con DCT y la comprimida con KLT. En sendos casos, la forma de onda es muy similar aun habiéndose llevado a cabo una compresión del 90%.

Las tablas 1 y 2 muestran los resultados obtenidos en términos de la SNR y de la métrica perceptual ODG (Objective Difference Grade) asociada a PEAQ. Como era de esperar, conforme el factor de compresión se incrementa, tanto la SNR como el parámetro ODG disminuyen. La SNR se calcula como:

$$SNR = 10\log_{10}\frac{\sum_{n=0}^{K-1} x^2(n)}{\sum_{n=0}^{K-1}(x(n)-\hat{x}(n))^2}.$$

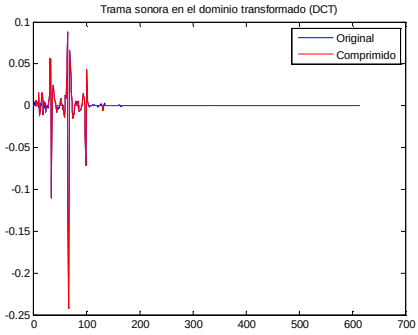

Fig. 3. Trama sonora en el dominio transformado (DCT). En color azul se ven los coeficientes de transformación originales y en rojo tras ser aplicada la retención por umbralización.

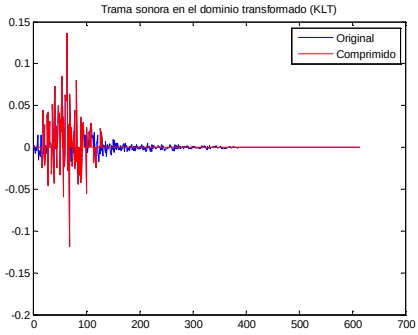

Fig. 4. Trama sonora en el dominio transformado (KLT). En color azul se ven los coeficientes de transformación originales y en rojo tras ser aplicada la retención por umbralización.

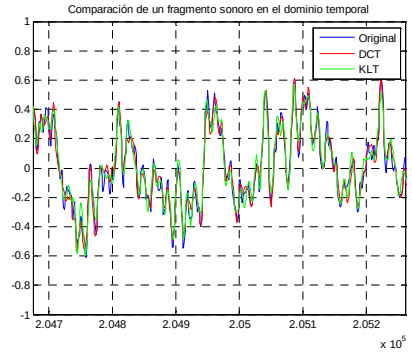

Fig. 5. Comparación del fragmento de audio original con los comprimidos a través del uso de la DCT (rojo) y la KLT (verde).

Las figuras 6 y 7 muestran las gráficas de los resultados recogidos en las tablas 1 y 2, respectivamente. El rendimiento en el uso de ambas transformadas es similar. Personalmente, no se aprecia degradación del fragmento sonoro hasta el 80% de compresión para el caso de la KLT y hasta el 90% de compresión para el caso de la DCT. Además, la degradación personalmente percibida es distinta en ambos casos. Para compresiones extremas, en el caso de la DCT se aprecia una notable pérdida de riqueza espectral, lo que se traduce en un sonido “apagado” y “tapado”. Sin embargo, en el caso de la

KLT, es más notorio un fondo ruidoso similar al que produciría un ruido blanco aditivo sobre la señal original. Sin embargo, personalmente creo que es factible emplear cualquiera de las dos transformadas combinadas con retención por umbralización para conseguir factores de compresión de hasta el 75% sin necesidad de combinarse con ninguna otra técnica, lo que equivaldría, por ejemplo, a que un fichero de audio *.wav* de 20MB pudiese reducir su volumen hasta 5MB sin apenas perder un ápice de su calidad inicial.

	SNR (dB)	
Nivel de Compresión	**DCT**	**KLT**
10%	77.1627	99.5660
20%	67.2080	90.2962
30%	60.2681	83.9384
40%	53.3141	54.2183
50%	43.6358	38.0084
60%	35.5652	30.2837
70%	28.7543	24.5700
80%	22.0034	19.2461
90%	14.1218	12.4743

Tab. 1. Comparativa de la SNR para la compresión mediante ambas transformadas en función del factor de compresión.

	ODG	
Nivel de Compresión	**DCT**	**KLT**
10%	0.188	0.195
20%	0.185	0.193
30%	0.173	0.192
40%	0.148	0.149
50%	0.051	-0.155
60%	-0.070	-0.867
70%	-0.689	-1.642
80%	-2.522	-2.700
90%	-3.648	-3.646

Tab. 2. Comparativa del ODG para la compresión mediante ambas transformadas en función del factor de compresión.

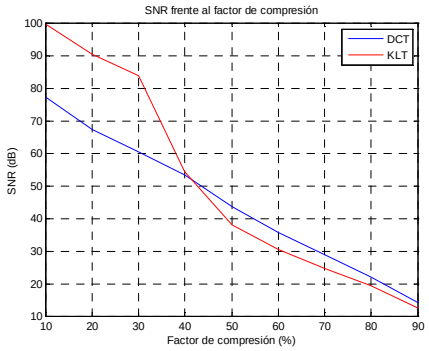

Fig. 6. SNR frente al factor de compresión para ambas transformadas.

Debido a que el rendimiento cuantitativo es similar, a que para factores altos de compresión no se aprecia el tipo de ruido comentado y a que el coste computacional es inferior, es preferible el uso de la DCT sobre la KLT para la compresión de la señal de audio o como parte de un sistema más complejo de codificación.

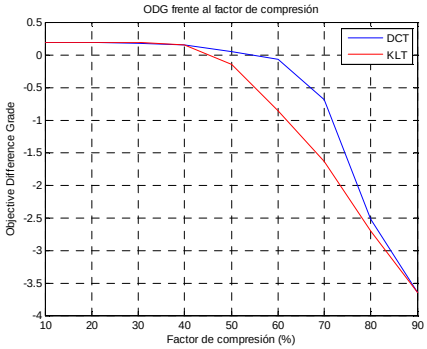

Fig. 7. ODG frente al factor de compresión para ambas transformadas.