Revista Facultad de Ingeniería
ISSN: 0717-1072
facing@uta.cl
Universidad de Tarapacá
Chile

Kaschel C., Héctor; Watkins, Francisco; San Juan U., Enrique

COMPRESIÓN DE VOZ MEDIANTE TÉCNICAS DIGITALES PARA EL PROCESAMIENTO DE SEÑALES Y APLICACIÓN DE FORMATOS DE COMPRESIÓN DE IMÁGENES

Revista Facultad de Ingeniería, vol. 13, núm. 3, 2005, pp. 4-10
Universidad de Tarapacá
Arica, Chile

Rev. Fac. Ing. - Univ. Tarapacá, vol. 13 Nº 3, 2005, pp. 4-10

COMPRESIÓN DE VOZ MEDIANTE TÉCNICAS DIGITALES PARA EL PROCESAMIENTO DE SEÑALES Y APLICACIÓN DE FORMATOS DE COMPRESIÓN DE IMÁGENES

Héctor Kaschel C.[1] Francisco Watkins[1] Enrique San Juan U.[1]

Recibido el 9 de marzo de 2005, aceptado el 14 de octubre de 2005

RESUMEN

El presente trabajo muestra la implementación de un modelo para compresión de voz a través del uso de técnicas digitales para el procesamiento de señales y de la aplicación de formatos de compresión de imágenes. Este trabajo se basa en la hipótesis que la compresión de la voz bajo este esquema logra una reducción significativa de la cantidad de *bytes* y de la consiguiente disminución de la velocidad en *bit/s* necesaria para la transmisión de la voz. La filosofía del modelo se centra en la conversión de tramas de voz en imágenes comprimidas, las que posteriormente son transmitidas a través del canal, para luego recuperar la trama de voz en el receptor a través de un proceso de síntesis. Para ello se emplean técnicas tanto en el dominio del tiempo como en la frecuencia, mediante la aplicación de filtros digitales IIR (*Infinite Impulse Response*), de la FFT (*Fast Fourier Transform*) y de variados formatos de compresión de imágenes.

Palabras clave: Compresión, Filtros IIR, FFT, JPEG, PNG, TIFF, MOS.

ABSTRACT

The present work shows the implementation of a model for compression of voice through the use of digital techniques for the signal processing and application of formats of compression of images. This work is based on the hypothesis that the compression of the voice under this scheme obtains a significant reduction of the amount of bytes and the consequent reduction of the necessary speed in bit/s for voice transmission. The philosophy of the model is centered in the conversion of frames of voice in compressed images, that later are transmitted through the channel, for later recovery, the frame of voice in the receiver, through a synthesis process. In order to obtain this, techniques in the dominion of time like in the frequency, are used by means of the application of IIR digital filters (Infinite Impulse Response), of FFT (Fast Fourier Transform), and of varied formats of image compression.

Keywords: Compression, Filters IIR, FFT, JPEG, PNG, TIFF, MOS.

INTRODUCCIÓN

Es improbable pensar en este tiempo la conveniencia de realizar transmisión de información multimedial en formato sin compresión. Principalmente, porque día a día son más las aplicaciones en este contexto que requieren de un amplio rango de calidad y *performance* de acuerdo a los requerimientos de usuarios heterogéneos. La alternativa es que sea posible la compresión masiva de los datos antes de efectuar su transmisión. Afortunadamente, un gran número de investigaciones durante las últimas décadas ha conducido a muchas técnicas y algoritmos de compresión que hacen factible la transmisión de multimedia.

Todos los sistemas de compresión requieren dos algoritmos: uno para la compresión de los datos en el origen y otro para la descompresión en el destino. En la literatura estos algoritmos se conocen como algoritmos de codificación y decodificación respectivamente. Para muchas aplicaciones un documento multimedia sólo se codificará una vez al almacenarse en el servidor, pero se puede decodificar miles de veces al ser vista por los clientes. Esta asimetría permite que el algoritmo de codificación sea lento y requiera *hardware* costoso, siempre y cuando el algoritmo de decodificación sea rápido y no requiera un *hardware* de alto costo. Por otra parte, para los multimedia de tiempo real, como las

[1] Departamento de Ingeniería Eléctrica, Facultad de Ingeniería. Departamento de Tecnologías Industriales, Facultad Tecnológica. Universidad de Santiago de Chile. Av. Ecuador 3519, Estación Central. Santiago-Chile. Fonos: (56) 2-7762963 - (56) 2–7786417 - (56) 2–6811100 A: 2396 Fax: (56) 2–6819079 hkaschel@lauca.usach.cl, fwatkins@lauca.usach.cl, esanjuan@lauca.usach.cl

videoconferencias y la voz sobre IP, la codificación lenta es inaceptable. Otra asimetría surge del hecho que no es necesario que el proceso de codificación/decodificación se pueda invertir. Por ejemplo, al comprimir, transmitir y descomprimir un archivo de datos el usuario espera recibir en forma correcta hasta el último *bit* de la información original. En multimedia este requisito no existe. Por lo general, es aceptable que la señal después de codificar y decodificar sea ligeramente diferente de la original. Los sistemas de codificación con pérdidas son importantes, porque aceptar una pequeña pérdida de información puede ofrecer ventajas enormes en la relación de compresión posible como, por ejemplo, el algoritmo de compresión de imágenes JPEG y el de compresión de voz LPC. [1-3].

FORMULACIÓN DEL MODELO

Conversión de voz en imagen

Esta técnica tiene como objetivo convertir tramas de voz en imágenes comprimidas, las que posteriormente son transmitidas, para luego realizar en el receptor el proceso de síntesis de modo de recuperar la trama de voz. Para conseguirlo se procesa la trama de voz a través de un banco de filtros pasabanda, los cuales eliminan redundancia en la información. El banco de filtros se encuentra dentro del rango de frecuencias en donde se ubican los principales formantes de la voz los que se detectan dinámicamente o son fijados en forma arbitraria (de acuerdo a estudios realizados por diversas investigaciones) de manera que las frecuencias en las que se localizan estos formantes determinen las frecuencias centrales de los filtros pasabanda. La señal filtrada por el banco de filtros es procesada posteriormente mediante el algoritmo FFT, como la transformada de *Fourier* entrega muestras complejas, la parte real e imaginaria de la señal filtrada en el dominio de la frecuencia, se convierten cada una por separado en una imagen equivalente a las que se les aplica un formato de compresión de imágenes como los conocidos JPEG, TIFF y PNG, comprimiendo de esta manera la trama de voz. Ambas imágenes son las que se transmiten y se reciben en el receptor, en donde se realiza el proceso inverso de manera de obtener la señal en el dominio del tiempo. El modelo descrito para las etapas de transmisión y recepción se muestran en las figuras 1, 2 y 3 respectivamente.

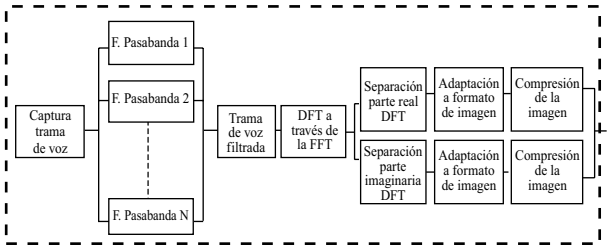

Fig. 1 Modelo de la etapa de compresión en el transmisor.

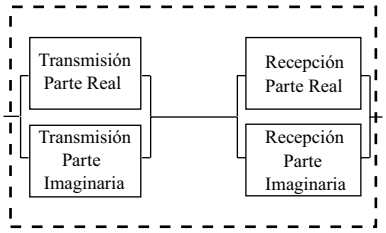

Fig. 2 Sistema de transmisión recepción.

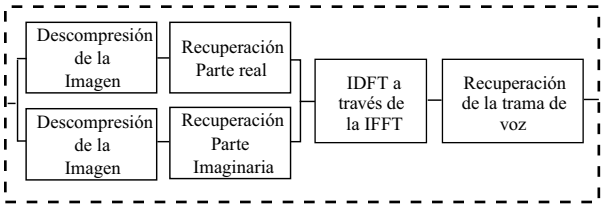

Fig. 3 Modelo de la etapa de descompresión en el receptor.

Estructura de un archivo WAV

En la tabla 1 se describe la estructura de un archivo WAV, cuyo conocimiento y manejo es clave para la implementación del programa de simulación del modelo de compresión planteado. Este formato es uno de los más utilizados para almacenar sonidos, se trata de almacenar las muestras una tras otra (a continuación de la cabecera del archivo, que entre otras cosas indica la frecuencia de muestreo), sin ningún tipo de compresión de datos, con cuantificación uniforme. La sencillez de este formato lo hace ideal para el tratamiento digital del sonido, de hecho Matlab tiene funciones para trabajar con este tipo de archivos [4].

Tabla 1 Formato de los archivos WAV.

Bytes	Contenido Usual	Propósito/Descripción
00-03	"RIFF"	Bloque de identificación (sin comillas).
04-07	Datos	Entero largo. Tamaño del fichero en bytes, incluyendo cabecera.
08-11	"WAVE"	Otro identificador.
12-15	"fmt "	Otro identificador
16-19	16, 0, 0, 0	Tamaño de la cabecera hasta este punto.
20-21	1, 0	Etiqueta de formato. .
22-23	1, 0	Número de canales (2 si es estéreo).
24-27	Datos	Frecuencia de muestreo (muestras/ segundo).
28-31	Datos	Número medio de bytes/segundo.
32-33	1, 0	Alineamiento de bloque.
34-35	8, 0	Número de Bits por muestra (normalmente 8, 16 ó 32).
36-39	Datos	Marcador que indica el comienzo de los datos de las muestras.
40-43	Datos	Número de bytes muestreados.
Resto	Datos	Muestras (cuantificación uniforme)

DESCRIPCIÓN DE LA IMPLEMENTACIÓN DEL MODELO PLANTEADO

En este apartado corresponde presentar la descripción para la implementación del modelo planteado presentado en las figuras 1, 2 y 3; la simulación se realiza mediante Matlab. También se muestra un diagrama de flujo asociado a dicha implementación (figura 4). En síntesis el modelo permite convertir una trama de voz en una imagen comprimida mediante los formatos JPEG, TIFF y PNG, y de esta forma realizar la transmisión a través de la red de una imagen comprimida que contiene la información de la voz, para realizar en el receptor el proceso de recuperación de la misma.

Las tramas de voz a procesar están en formato WAV. Las funciones que realiza el programa de simulación que implementa el modelo de las figuras 1, 2 y 3 son las siguientes:

Selección del sonido

Esta opción permite seleccionar las tramas de voz a procesar, las que como ya se ha mencionado se encuentran en formato WAV.

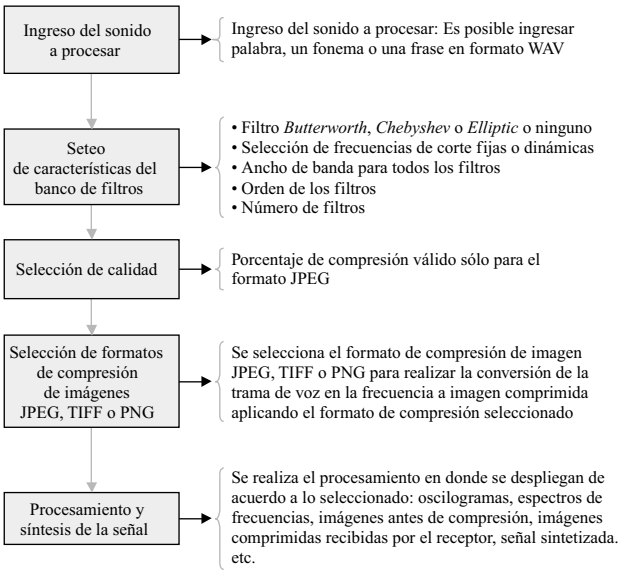

Fig. 4 Diagrama de flujo para la simulación del modelo.

Selección de banco de filtros

Los bancos de filtros a seleccionar son de dos tipos: fijos y dinámicos. En los filtros fijos las ubicaciones de las frecuencias de cortes y anchos de banda para cada uno de ellos se determinan arbitrariamente de acuerdo a la ubicación de las dos primeras formantes para las cinco vocales españolas. En relación con este concepto se muestra una tabla para los cuatro filtros pasabandas fijos que se incorporan en el programa, que cubren las ubicaciones de las formantes especificadas en la tabla 2. El banco de filtros dinámicos consiste en determinar dinámicamente las frecuencias de corte de una cantidad de filtros definidos por el usuario, los que pueden ser *Butterworth*, *Chebyshev* o *Elliptic* dentro de un rango entre 0 y 4 kHz; además, se debe, seleccionar el orden de los filtros y el ancho de banda que tendrán, el cual es común a todos ellos. La metodología para seleccionar la ubicación de los filtros dentro del rango 0 a 4 kHz se centra en un análisis del espectro de frecuencias de la trama completa de la señal de voz, mediante la determinación de los *peaks* de amplitud del espectro. Se observa que la multiplicación entre el ancho de banda seleccionado para los filtros y la cantidad de filtros seleccionada no supera los 4 KHZ, es decir, que cumple la siguiente relación. N^o *de filtros x BW* $\leq$ *4 kHz*. El programa no considera otro *peak* dentro de un mismo ancho de banda en donde ya se haya encontrado un *peak* [5].

Tabla 2 Frecuencias de corte para los filtros pasabandas fijos que cubren las 2 primeras formantes de las vocales españolas.

Filtro Nº	Frecuencia de corte inferior (Hz)	Frecuencia de corte superior (Hz)	Ancho de banda (*BW*)
1	200	300	100
2	500	800	300
3	1150	1500	350
4	2000	2500	500

De esta forma, el programa encuentra la cantidad de filtros seleccionados con un determinado BW que es el mismo para todos los filtros. Otro parámetro importante es el orden del filtro el cual es seleccionable tanto para los fijos como para los dinámicos. La trama de voz es procesada por cada filtro mediante convolución en el tiempo; de esta manera se consigue un importante ahorro de tiempo respecto a que si la aplicación de los filtros fuera en el dominio de la frecuencia.

Reducción de redundancia mediante banco de filtros

La aplicación del banco de filtros es fundamental, ya que se consigue quitar redundancia a la señal, esto permite además que para muchas frecuencias los niveles de amplitud del espectro sean muy pequeños o cero, lo que significa que al convertir una trama de voz filtrada en una imagen, estos niveles se presenten como niveles de grises muy similares, lográndose una mayor compresión de la información al aplicar el formato de compresión de imágenes seleccionado.

Conversión de la señal de voz en imagen

Al aplicar el algoritmo FFT a la señal filtrada por el banco M de filtros pasa banda, se obtienen N muestras complejas conjugadas, por lo que para propósitos de transmisión basta sólo con procesar una de las muestras de cada par conjugado, es decir, un total N/2 muestras, produciéndose de esta forma un importante ahorro de BW del canal. Luego, es conveniente procesar la parte real e imaginaria de las muestras por separado. Así, N/2 números reales y N/2 imaginarios se convierten a formato de imagen en escala de grises respectivamente, obteniéndose de esta manera un formato de imagen de la parte real y otra de la parte imaginaria.

Aplicación de formatos de compresión de imágenes

Al obtener la información de la trama de voz en la frecuencia en dos vectores con formato de imagen, es posible aplicarles algún formato de compresión de imagen como el JPEG, que por su probada buena relación calidad/compresión constituye una buena alternativa para el modelo planteado, además que permite decidir la calidad de la compresión a usar. Para poder realizar una medición cualitativa y cuantitativa de la eficacia del formato JPEG en este modelo, resulta interesante compararlo con respecto a otros formatos. Por esta razón es que se incorporan también los formatos TIFF y PNG. Como se aprecia en la figura 1, el formato de compresión es aplicado por separado a la parte real e imaginaria de la DFT, respectivamente llevadas a formato de imagen, creándose de esta forma dos imágenes comprimidas, las que son transmitidas a través del canal. [6-11].

Síntesis de voz en el receptor

Las imágenes comprimidas de las partes reales e imaginarias de la DFT llegadas al receptor se convierten al dominio de la frecuencia, obteniéndose de esta manera N/2 muestras complejas del tipo *a+jb*, la conversión de la imagen comprimida de la parte real nos da la componente real *a* y la conversión de la imagen comprimida de la parte imaginaria nos da la parte imaginaria *b*, luego se deben construir sus conjugados (*a-jb*), para de esta forma tener las N muestras complejas constituyentes de la DFT. Si se denomina DFT_T a la DFT de la señal filtrada por el banco de filtros en el transmisor y a la DFT en el receptor como DFT_R, se observa que ambas no son iguales, debido a que la DFT_R proviene de un proceso de compresión de imágenes el cual aporta pérdidas. El proceso de síntesis se logra al aplicar la IDFT a la DFT_R, luego de esta forma se obtiene la trama de voz sintetizada [12].

DESARROLLO DE LA SIMULACIÓN

A continuación se presenta un ejemplo para una simulación en particular; en este caso se procesará una frase hablada por una mujer, que dice: "El perro olfatea la comida de su amo", la cual tiene una duración de 3.573ms, cuyas muestras procesadas en PCM tienen un peso de 28,485 *kbytes*. Los datos introducidos en la interfaz gráfica para la simulación se dan en la tabla Nº 3. [13-15].

Tabla 3 Ajustes para una Simulación (continúa)

Ajustes	Selección
Calidad de compresión	50
Tipo de filtro para el banco: Butteworth, Chebyshev, Elíptico o ninguno	Butterworth
Selección de banco de filtros: fijos o dinámicos	Fijos

Tabla 3 Ajustes para una Simulación (continuación)

Ajustes	Selección
Ancho de banda BW para todos los filtros del banco, 200, 300, 400, 500 o según tabla 2	Según tabla 2
Orden del filtro 2 ó 4	4
Número de filtros componentes del banco	Según tabla 2
Nombre o tipo de archivo wav a procesar	Frase
Selección del formato de compresión: JPEG, TIFF o PNG	JPEG

En la figura 5 se muestra el oscilograma o gráfico temporal de la señal original antes de ser procesada y en la figura 6, el espectro de frecuencias imaginario luego de aplicar la FFT a la salida del banco de filtros; en ambos gráficos los espectros se encuentran dentro de un rango comprimido y normalizado entre 0 y 1 para su posterior conversión en imagen.

En la figura 7 se muestran las imágenes equivalentes para estos espectros antes de su compresión en el transmisor.

En la figura 8 se puede apreciar las imágenes reales e imaginarias en el receptor descomprimidas.

En la figura 9 se muestra la voz sintetizada a partir de la conversión de estas imágenes a las partes real e imaginaria de la FFT en el receptor y luego calcular la IFFT.

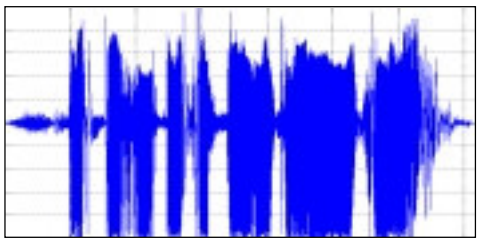

Fig. 5 Señal original en el tiempo antes de procesar.

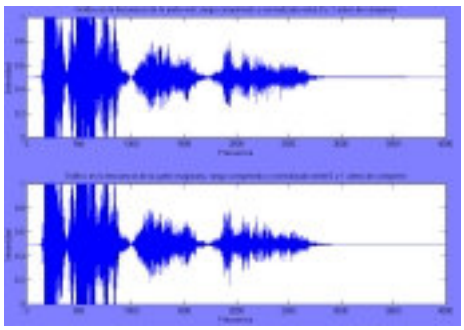

Fig. 6 Espectros de frecuencias reales e imaginarios.

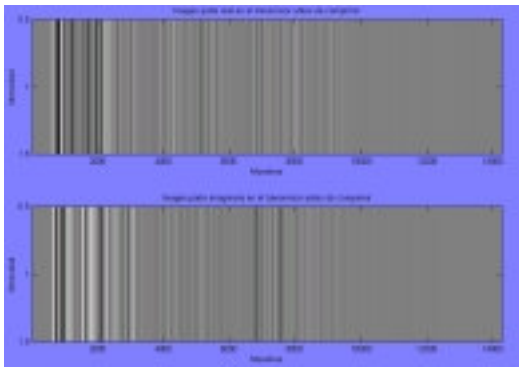

Fig. 7 Imágenes real e imaginaria en el transmisor antes de comprimir.

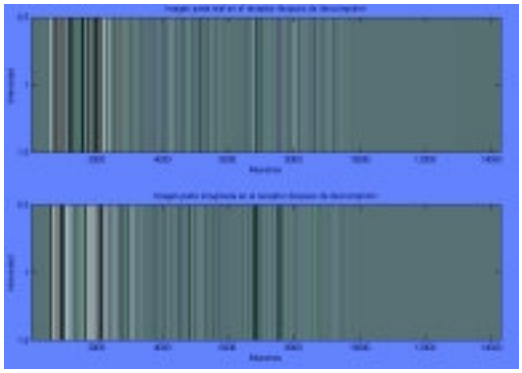

Fig. 8 Imágenes real e imaginaria en el receptor después de descomprimir.

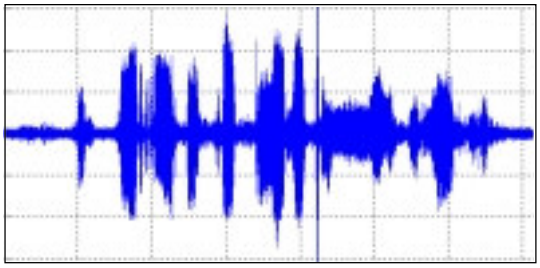

Fig. 9 Oscilograma de la señal sintetizada en el receptor.

Desempeño de los filtros

En relación con la *performance* del banco de filtros pasa-banda, es evidente que la disminución en el número de filtros que componen el banco favorece la compresión, pero se produce, sin embargo, una diferencia en la calidad audible de la voz. En este caso la diferencia no se debe principalmente a ruido, sino que debido a la aplicación de distintos números de filtros que componen el banco, lo cual trae como consecuencia una alteración paulatina en las características de la voz del hablante a

medida que disminuye el número filtros que componen el banco. El caso extremo se encuentra cuando el banco está constituido sólo por un filtro pasabanda, en donde un ancho de banda del filtro inferior a 400Hz hace ininteligible la voz sintetizada. El hecho de aplicar diferentes números de filtros al banco de filtros dependerá de las características del ensayo en particular y de los niveles de compresión que se deseen obtener. La aplicación del banco de filtros aporta hasta un 60% más de compresión respecto a la no aplicación del banco de filtros. Respecto a cuál es el filtro más adecuado para emplear en el modelo, se obtiene que las diferencias entre los tres tipos de banco de filtros son muy sutiles y los resultados encontrados señalan que el banco de filtros *Butterworth* posee el mejor comportamiento en lo que respecta a compresión; en cuanto al banco de filtros *Elliptic* es el que presenta inferior desempeño. Como las diferencias son mínimas, se recomienda seleccionar el de *Butterworth* que es el más común y simple de implementar. Es importante realizar algunos comentarios acerca de los bancos de filtros dinámicos y fijos: El banco de filtros fijos logra en general un nivel de compresión levemente inferior al banco de filtros dinámicos, ya que el banco de filtros fijos cubre un mayor rango de frecuencias respecto de los dinámicos. Lo que implica, además, obtener una calidad audible levemente superior respecto a si el proceso se realizara con el banco de filtros dinámicos.

Desempeño de los formatos de compresión de imagen

Respecto de qué formato de compresión de imagen resulta más adecuado o efectivo, dependerá evidentemente de la aplicación particular que se requiera realizar. A base de los resultados obtenidos, los desempeños más efectivos son los logrados con los formatos JPEG y PNG, y luego el formato TIFF. El formato JPEG posee la ventaja de poder seleccionar la calidad del porcentaje de compresión, lo que permite una gran flexibilidad convirtiéndolo en un formato muy conveniente para este modelo de compresión de voz y, además, que permite trabajar hasta con calidades que llegan a un 1% y aún así en algunos casos obtener una señal pobre en calidad, pero inteligible. Con este formato incluso hasta con un 25% de calidad puede obtenerse una voz sintetizada con calificación MOS 4,5 lográndose en este caso compresiones de imágenes para transmisión comprendidas entre aproximadamente un 54 y un 71%, dependiendo del número de filtros usados para el banco. Para el formato PNG los porcentajes de compresión en general son comparables con JPEG entre un 25 y un 50% de calidad. En lo relacionado con el MOS, PNG es comparable con JPEG en el rango comprendido entre un 75 y un 25% de calidad. Estas características convierten a PNG en un formato altamente conveniente de emplear al igual que JPEG, a pesar de no tener la posibilidad de seleccionar la calidad de compresión. Por último, el formato que aporta menor compresión es el TIFF, con cerca de un 50% de lo obtenido por el formato PNG [16].

CONCLUSIONES

Tras la evaluación de los resultados, se puede concluir que el objetivo de este trabajo resulta satisfactorio, ya que es posible convertir tramas de voz en imágenes, las cuales son comprimidas mediante un formato de compresión de imágenes para posteriormente ser transmitidas a través del canal y luego recuperar la trama de voz en el receptor a través de un proceso de síntesis, cumpliéndose de esta forma la hipótesis planteada que postula que la compresión de la voz bajo este esquema logra una reducción significativa de la cantidad de *bytes* y de la consiguiente disminución de la velocidad en *bit/s* necesaria para la transmisión de la información. Se ha comprobado que el modelo de compresión de voz formulado en este trabajo es viable y es posible lograr compresiones importantes de la voz, pudiéndose alcanzar en algunos casos valores de alrededor de un 60% de compresión (MOS 4,5, banco de filtros fijos) con una calidad razonable dependiendo de la aplicación en particular. Los valores de compresión obtenidos son variados y dependen en gran medida del "seteo" de filtros, de los formatos de compresión seleccionados y del nivel de calidad audible que se desee; sin embargo, para una calidad de MOS entre 4,0 y 4,7 aproximadamente se pueden obtener velocidades entre 15,5*kbit/s* y 42,7*kbit/s* respectivamente El modelo presentado puede resultar innovador, ya que a primera vista no pareciera evidente la conversión de tramas de voz en imágenes equivalentes. Bajo este esquema será posible transmitir las imágenes de audio junto con las imágenes de video propiamente tal, ambas por un mismo canal en forma apropiada evitándose de esta forma probablemente problemas de sincronización entre audio e imagen, con el consiguiente ahorro de un canal y mejoramiento del servicio. Se podrían seguir dando otras posibles aplicaciones de este modelo, dependiendo en gran medida de las necesidades particulares que se presenten. Cabe destacar que el modelo proporciona muy buenas expectativas en lo que respecta a compresión de la voz, por lo que resulta interesante seguir considerándolo en aplicaciones experimentales y a futuro constituir una buena alternativa de esquema de compresión de voz en aplicaciones reales.

REFERENCIAS

[1] A.S. Tanenbaum. "Computer Networks". Pearson. 3rd edition. 1997.

[2] S. Keagy et al. "Integration Voice and Date Networks". Cisco System. Pearson. 2001.

[3] J. Davidson, J. Peters. "Voice over IP Fundamentals". Cisco System. Pearson. 2001.

[4] "Formato de los ficheros wav". http://www.upv.es/protel/usr/jotrofer/sonido/sound.htm

[5] M. Faúdez Zanuy. "Tratamiento digital de voz e imagen". Marcombo. 2000.

[6] A. de la Escalera. "Visión por computador". Prentice Hall. 2001.

[7] K.R. Castleman. "Digital Image Processing". Prentice-Hall. Englewood Cliffs. New Jersey, 07632. 1996.

[8] B.E. Usevitch. "A tutorial on modern lossy wavelet image compression: foundations of JPEG 2000". IEEE Signal Processing Magazine. Vol. 18. Issue 5, pp. 22-35. September. 2001.

[9] A. Skodras, C. Christopoulos, T. Ebrahimi. "The JPEG 2000 still image compression standard". IEEE Signal Processing Magazine. Vol. 18. Issue 5, pp. 36-58. September. 2001.

[10] D. Olivo. "JPEG: Metodo per la compressione di immagini". Maggio 1998. http://utenti.lycos.it/debolivo/jpeg/indexj.html

[11] "JPEG". http://coco.ccu.uniovi.es/immed/compresion/descripcion/jpeg/jpeg.htm

[12] J.G. Proakis and D.G. Manolakis. "Digital Signal Processing. Principles, Algorithms and Applications". Prentice Hall, INC. 1998.

[13] V.K. Ingle, J.G. Proakis. "Digital Signal Processing. Using Matlab V.4". PWS Publishing Company. 1997.

[14] Maurice Bellanger. "Digital Processing of Signals. Theory and Practice". John Wiley & Sons. 1986.

[15] "Signal Processing Toolbox. For use with Matlab". User's Guide version 4.2. The MathWorks Inc. 1999. http://www.mathworks.com

[16] L.A. Ferreres, M.J. Roca Estellés. "Sistemas de difusión de información audiovisual en Internet: Formato Gráfico PNG". http://www.conganat.org/iicongreso/comunic/008/png.htm