

# Sonificación de Imágenes 2D

Antonio S. MONTEMAYOR, Alberto L. CORRALES, Ángel SÁNCHEZ  
Dpto. Informática, Estadística y Telemática, Universidad Rey Juan Carlos  
Móstoles, Madrid 28933, España

## RESUMEN

La transformación de información de un medio a otro ha sido, y sigue siendo, un tema de investigación activo. Dentro de este campo podemos incluir la sonificación, que se define como la transformación de relaciones de datos en relaciones acústicas con el propósito de facilitar la comunicación y la interpretación. En los últimos años y, gracias a las facilidades multimedia que han aportado las nuevas tecnologías, la sonificación de datos se ha potenciado en gran manera. El acceso a la información de las personas con discapacidades visuales se ha visto favorecido por este motivo, aunque aún hay grandes barreras. Actualmente existen numerosas aplicaciones de lectura de textos en documentos o páginas web basados en sistemas OCR, así como del paso inverso, conversión de voz a texto. Sin embargo para la percepción y el entendimiento de las imágenes aún no hay buenas soluciones. El optófono es el instrumento capaz de sonificar una imagen o escena. En este trabajo se presenta un prototipo para el que se han experimentado diversas y muy conocidas técnicas de segmentación de imágenes.

**Palabras Clave:** sonificación, procesamiento de imágenes, segmentación, optófono, tiflotecnía.

## 1. INTRODUCCIÓN

La transformación de información de un medio a otro puede resultar de especial interés en casos en los que es difícil asimilar una determinada cantidad o ciertos tipos de datos. El uso de sonido no hablado (nonspeech audio) para tratar o canalizar información o más específicamente, la transformación de relaciones de datos en relaciones acústicas con el propósito de facilitar la comunicación y la interpretación, se denomina sonificación [9]. Bajo este nombre, se engloban muchas técnicas multidisciplinares, que integran conceptos de la percepción humana, la acústica, las artes o el diseño, por citar algunos, y que tiene cabida a la acción combinada de profesionales de la informática, la física, la música o la psicología entre otros. Gran parte de la investigación centrada en la sonificación se ha dirigido a sentar las bases teóricas de los umbrales auditivos, las escalas psicofísicas y los modelos de percepción auditivos. En particular, sobre los parámetros involucrados, como son la intensidad, la frecuencia y la discriminación temporal de sonidos estáticos [5], [11], [12], los determinantes de la tonalidad y volumen, los efectos de enmascaramiento [4], y propiedades de localización auditiva [1]. El tratamiento de grandes cantidades de información mediante la incorporación del sonido es un campo de investigación de innegable utilidad. El sonido es esencialmente una onda de presión propagada sobre un medio material. Las características del sonido son el volumen (directamente relacionado con la amplitud de la onda), el tono (relacionado logarítmicamente con la frecuencia de la onda), el timbre, la localización, la duración y el ritmo [9]. Generar una nueva dimensión como es el sonido, haciendo

variar sus características, para la visualización de los datos nos permite aprovechar de forma paralela las relaciones entre ellos [6].

Dentro de la sonificación podemos incluir una interesante aplicación: el optófono. El optófono es el instrumento capaz de sonificar una imagen basando su sonido en características de ésta. Este instrumento puede tener especial relevancia en personas con alguna discapacidad visual, pues en algunos casos la percepción del entorno puede verse limitada a canales de recepción como el oído y el tacto.

Para el presente trabajo hemos desarrollado en software un optófono similar al del investigador Peter Meijer [10], que tiene en cuenta la intensidad de los niveles de gris y la posición de éstos en la imagen en escala de grises para realizar la sonificación. La finalidad de este prototipo es la de experimentar con diferentes métodos de segmentación de imágenes que resulten más apropiados para la sonificación. Asimismo, nos sirve como paso preliminar para un trabajo más ambicioso como explicaremos en la última sección. Hay que entender que una de las limitaciones más acusadas de la transferencia de información a través del sonido es la restricción de memoria que tiene una persona frente a un conjunto denso o continuo de información de este tipo. La sonificación de una imagen compuesta por muchos detalles no permitirá al usuario discernir lo superfluo de lo trascendente, restando funcionalidad al instrumento. Por este motivo, es de vital importancia la simplificación de la escena a sonificar.

A continuación, se describe la estructura del resto del trabajo. En la sección 2 se detalla la problemática de la sonificación de imágenes, y en concreto se presenta el prototipo de optófono implementado. En la sección 3, se discuten los aspectos derivados de la simplificación de las imágenes, describiendo los métodos de segmentación utilizados en este trabajo. La sección 4 se presentan algunos resultados experimentales y finalmente en la sección 5 se muestran las conclusiones del presente trabajo y se reflexiona sobre futuros avances.

## 2. SONIFICACIÓN DE IMÁGENES

Nuestro optófono genera sonido a partir de una imagen digital en escala de grises en base a las características visuales de la misma. Para ser más precisos, sonificamos una imagen por columnas, creando una onda sonora de cada uno de los píxeles que contiene. Una intuitiva y estrecha correspondencia entre las características de la imagen y el sonido a producir es el utilizar el nivel de intensidad de cada píxel para dar mayor o menor volumen al sonido generado y utilizar su posición espacial en la dirección vertical para determinar el tono, eligiendo frecuencias mayores para los píxeles a mayor altura en la imagen y viceversa. Para cada columna sumamos las ondas generadas por cada uno de sus píxeles y hacemos un barrido de izquierda a derecha sonificando así todas las columnas y por tanto toda la imagen (Fig. 1).

De esta explicación se deduce que el nivel de gris de cada píxel está directamente relacionado con la amplitud de la onda que genera (el volumen del sonido). Asimismo, la altura de cada píxel determina la frecuencia de la onda (que está relacionada íntimamente con el tono del sonido), siendo mayor cuanto más alto se encuentra en la columna.

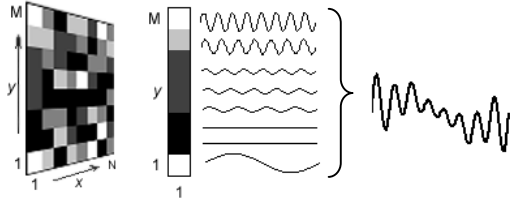


Fig. 1: Sonificación de la primera columna de una imagen ficticia a partir de sus píxeles constituyentes. La suma de los sonidos generados para cada píxel es el sonido resultante para la columna.

La función sinusoidal utilizada para la generación de la onda correspondiente a la sonificación de cada píxel viene dada por la Ec. (1).

$$S(t) = A(I(y)) \cdot \text{sen}(\omega(y) \cdot t) \quad (1)$$

siendo  $S$  el valor de la onda en cada instante  $t$ ,  $A$  su amplitud, dependiente del nivel de intensidad del píxel en la columna,  $I(y)$ , y  $\omega$  su frecuencia, dependiente de la altura del píxel en la columna,  $y$ .

Por poner un ejemplo, en Fig. 1 se observa que la onda correspondiente al píxel blanco de la fila más baja tiene la misma amplitud que su homóloga de la fila superior, sin embargo ambas difieren en su frecuencia. La suma de la contribución de cada uno de los píxeles es el sonido correspondiente a la primera columna. Para sonificar la imagen entera, barremos las sucesivas columnas, generando así el sonido total compuesto por los sonidos individuales de cada una de ellas. La contribución a la onda de la columna de los dos píxeles blancos del sencillo ejemplo viene expresado por la Ec. (2) y Ec. (3) respectivamente.

$$S_1(t) = A(255) \cdot \text{sen}(\omega(8) \cdot t) \quad (2)$$

$$S_2(t) = A(255) \cdot \text{sen}(\omega(1) \cdot t) \quad (3)$$

siendo 255 el nivel de gris correspondiente al color blanco y 1 y 8 las alturas respectivas de ambos píxeles. También es interesante observar en la misma figura que los píxeles negros generan ondas de amplitud igual a cero, sean cuales sean sus alturas, y por tanto sus frecuencias.

Entonces, la función resultante para una columna  $x$  de la imagen se expresa a través de la suma de las contribuciones dadas por la Ec. (1) de cada uno de los píxeles que contiene, llegando así a la Ec. (4).

$$S_T(x, t) = \sum_{y=1}^M A(I_{x,y}) \cdot \text{sen}(\omega_y \cdot t) \quad (4)$$

siendo  $x=1..N$  la posición de la columna en la imagen de dimensiones  $N \times M$ ,  $y=1..M$  la posición del píxel en la columna,  $I_{x,y}$  el nivel de gris del píxel situado en la posición  $(x, y)$  de la imagen, y  $\omega_y$  la frecuencia perteneciente a la altura  $y$ . La adhesión sucesiva de todas las funciones suma correspondientes a las columnas formará finalmente el sonido total de la imagen. De esta manera, la duración del sonido debido a la sonificación de toda la imagen será suma de las duraciones parciales de las ondas de todas las columnas.

Es fácil darse cuenta de que el acoplamiento de las funciones correspondientes a las columnas para crear la sonificación de toda la imagen creará sonidos extremadamente complejos y sin aparente sentido. Esta supuesta complejidad se incrementa notablemente con imágenes grandes, compuestas por multitud de niveles de gris y con detalles finos. De aquí se deduce la alta necesidad de preprocesar la imagen original y simplificarla en base a las características generadoras del sonido, es decir, el número de niveles de gris de sus píxeles y el tamaño de la misma. En la siguiente sección abordaremos los diversos tratamientos que hemos elegido para realizar nuestros experimentos.

El sonido creado tras el barrido de la totalidad de la imagen representará de una manera sonora a la propia imagen y llevará toda la información que en principio pudiera contener en su representación habitual. Por este motivo, y aunque “escuchar” una imagen no sea una tarea sencilla, podemos llegar a entenderla de una forma grosera. Esta aproximación al significado de la escena puede tener especial importancia en personas con discapacidades visuales, y en concreto en casos de ceguera parcial o total.

### 3. PROCESAMIENTO DE IMÁGENES

Hemos visto en la sección anterior que las imágenes con exceso de detalle, con múltiples niveles de gris o de gran tamaño pueden generar sumas de ondas cuya sonificación es potencialmente compleja. Es por tanto necesaria una labor de simplificación que llevaremos a cabo en una etapa de preprocesamiento.

#### Umbralización

Una manera de disminuir drásticamente el número de niveles de gris de una imagen es aplicando una umbralización. Mediante una umbralización simple y global dividimos el histograma de la imagen usando un umbral único y provocando la binarización de la misma. De esta manera pasamos a tener en la imagen tan solo dos niveles de intensidad, blanco y negro en base, al umbral aplicado y a los valores de intensidad de los píxeles. En la práctica, este método sólo será aplicable para imágenes tomadas en entornos altamente controlados [4]. En efecto, aplicar la umbralización puede provocar un exceso de píxeles blancos o negros dependiendo de los niveles de gris originarios en la imagen y el umbral escogido (Fig. 2). Por esta razón, es preferible utilizar la umbralización junto con algún método añadido de segmentación.

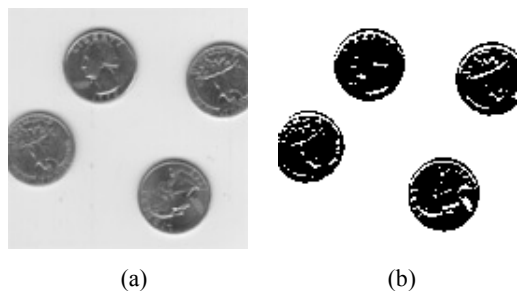


Fig. 2: a) imagen original; b) imagen umbralizada

### Detección de Bordes

La detección de bordes en imágenes en escala de grises es una de las técnicas de segmentación más utilizada y, con mucho, la aproximación más común para detectar discontinuidades importantes [3]. Consiste en destacar los píxeles correspondientes a las fronteras entre regiones de niveles de gris diferentes. Por esta razón, se consigue una reducción en el número de píxeles principales, pues pasan a ser tan sólo los ubicados en las fronteras de regiones homogéneas (Fig. 3).

Así pues, la idea subyacente en la mayoría de las técnicas de detección de bordes es el cálculo de un operador de derivada local. En muchos casos se adoptan aproximaciones al operador gradiente a través de convoluciones con máscaras de convolución determinadas. Ejemplos de estas máscaras son, entre otras, las de Sobel, de Prewitt o de Roberts. En nuestro prototipo hemos escogido la de Sobel por ser ampliamente utilizada en la bibliografía.

En nuestro prototipo hemos incluido filtros de suavizado que normalmente se utilizan para evitar una sobredetección causada por el ruido o imperfecciones de la imagen. La aplicación de estos filtros es opcional aunque recomendada. En concreto, los filtros de suavizado implementados son los de media local y gaussiana.

Una cualidad que puede interesar a la hora de sonificar una imagen es la de distinguir las siluetas de los objetos presentes en la escena. Esta interpretación se puede conseguir con la técnica de detección de bordes sin perder generalidad. Sin embargo, si resaltamos los bordes es preferible umbralizar la imagen de bordes resultante para simplificar el futuro sonido resultante tras la sonificación. De esta manera, evitamos introducir niveles de amplitud muy diferentes acotándolos a tan solo dos.

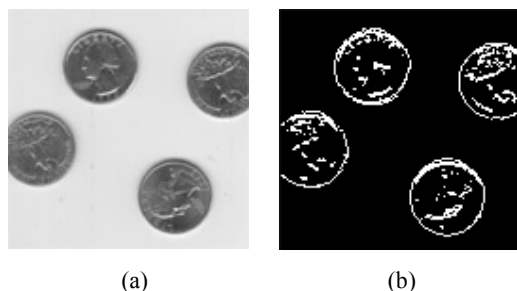


Fig. 3: a) Imagen original; b) imagen de bordes correspondiente

### Reducción de Resolución Espacial

Una forma directa y efectiva de disminuir la cantidad de información a procesar es mediante una reducción espacial de la imagen original. El número de píxeles disminuye y por tanto el número de sonidos a crear también. En la reducción del tamaño de una imagen el factor clave a tener en cuenta es el método de interpolación que utilizamos para que la imagen no pierda contenido semántico. Se trata de reducir el número de píxeles manteniendo las proporciones originales. Los métodos de interpolación más simples y utilizados son los de interpolación bilineal, bicúbica y por vecindad. Los dos primeros tienen en cuenta los niveles de los píxeles próximos para alcanzar un compromiso con el que se recalcula, generando un nuevo valor diferente en general de los anteriores. La interpolación por vecindad es más simple y no añade nuevos valores de intensidad, sino que utiliza los pertenecientes a la imagen. Por esta razón elegimos esta técnica en nuestro prototipo, simplificando aún más la imagen resultante aunque siempre manteniendo un cierto compromiso visual respecto a la original.

### Cuantización de Niveles de Intensidad

Las pequeñas diferencias en amplitud de un sonido se perciben peor que las otras características como por ejemplo las del tono [9]. Reducir el número de posibles amplitudes hará aumentar las diferencias entre ellas y las posibilidades de discernirlas mejor. Observando la Ec. (1) notamos que el número de posibles amplitudes está ligado íntimamente con el número de niveles de gris en la imagen. Deducimos entonces que la reducción del número de niveles de gris de la imagen implica una simplificación deseable en el resultado final de la sonificación.

Para que las reducciones de niveles de intensidad tengan algún efecto notable en el sonido generado por sus píxeles es deseable que el número de niveles de gris de la imagen final no sea mayor de 16, incluso menos si ésta se compone de finos detalles (Fig. 4).

## 4. RESULTADOS EXPERIMENTALES

Ya se ha mencionado que adoptamos una correspondencia entre la altura del pixel a sonificar y la frecuencia de la onda a generar. Esta correspondencia se basa en una escala de frecuencias que debemos elegir. En nuestro prototipo hemos implementado varias escalas ya utilizadas en el trabajo del profesor Peter Meijer, como la escala lineal, exponencial, la de Mel o la de Bark [10]. Sin embargo, hemos creído oportuno añadir entre las opciones de sonificación la escala natural musical, puesto que puede expresar de modo más intuitivo la geometría expuesta por los píxeles de la imagen. Nuestra hipótesis es que una persona occidental tiene el oído culturalmente educado de tal manera que dos notas musicales consecutivas que no guardan la misma distancia en frecuencia que otro par consecutivo, lo interpreta como equidistante. Este hecho es fácilmente demostrable. Definimos un salto unitario como el intervalo entre una nota musical y la siguiente. De ésta forma, a pesar de que los intervalos Mi4-Fa4 y Fa4-Sol4 son de un semitono y un tono respectivamente, según nuestra definición corresponderían ambos a un salto unitario (Fig 5b). Para este ejemplo, las distancias en frecuencia de ambos intervalos son de 20 y 43 Hz, lo que contrasta con las demás escalas de frecuencias utilizadas, principalmente la lineal (Fig. 5a) pues queda patente que la escala natural es definitivamente no lineal.

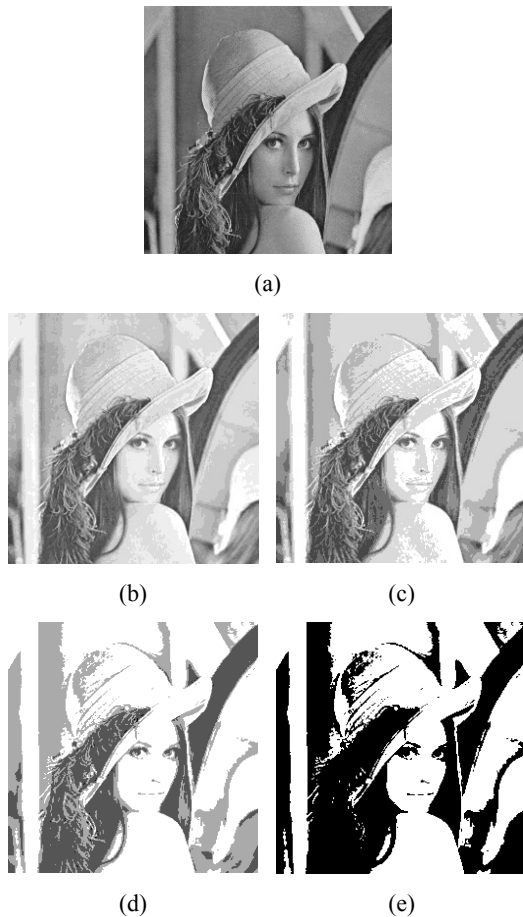


Fig. 4: a) Imagen original, 256 niveles de gris; b) imagen cuantizada a 16 niveles de gris; c) a 8 niveles de gris; d) a 4 niveles de gris; e) a 2 niveles de gris

La escala natural la utilizamos fundamentalmente para imágenes de dimensiones pequeñas pues tratamos de no utilizar más notas que las que contenga una octava. Sin embargo, otra escala que podemos utilizar es la cromática para la que nuestro salto unitario se corresponde con un semitono. Esta escala es útil para imágenes algo más grandes pues en una única octava tenemos hasta 12 sonidos diferentes. (Fig 5c)

En primera aproximación y para comprobar la bondad de la sonificación hemos hecho pruebas con nuestro prototipo y con imágenes sintéticas generadas externamente. En Fig.6 se muestran dos imágenes que resultan especialmente útiles para experimentar con las escalas de frecuencias y con la dependencia de la amplitud de las ondas con el nivel de gris del píxel sonificado respectivamente. Una diagonal cuyos píxeles son de igual nivel de intensidad en un fondo negro servirá para discernir posibles diferencias entre las escalas de frecuencias utilizadas para sonificar cada uno de sus píxeles (Fig. 6a). Una recta horizontal cuyos píxeles tengan niveles de gris variables, de oscuros a claros, nos mostrará la dependencia de las amplitudes de las ondas generadas con los niveles de intensidad de sus píxeles manteniendo fijas las frecuencias (Fig. 6b).

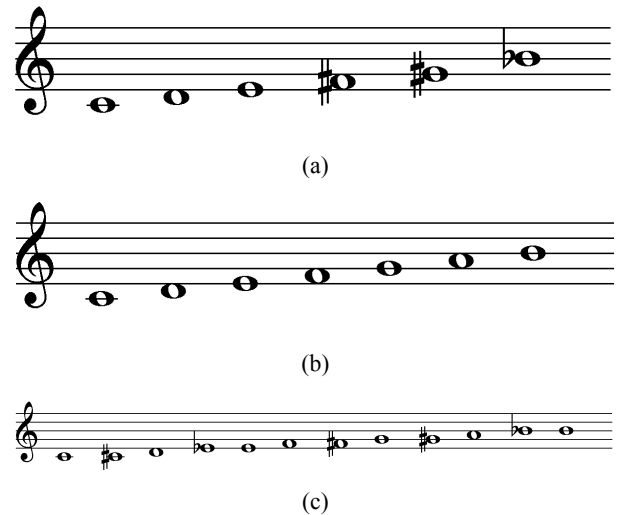


Fig. 5: a) Escala de tonos completos; b) escala natural; c) escala cromática

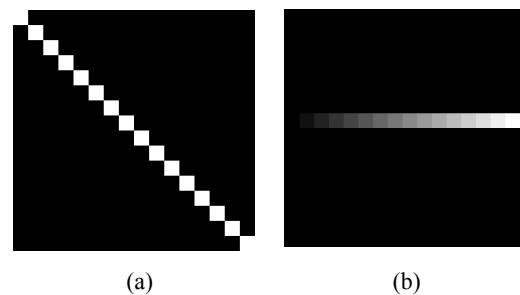


Fig. 6: a) Diagonal principal con píxeles blancos en fondo negro; b) recta horizontal con píxeles en escala de grises en fondo negro

Para imágenes más realistas empezamos preprocesándolas con los métodos descritos en la sección anterior. En concreto, la detección de bordes es muy determinante en la extracción de las propiedades geométricas de los objetos presentes en la escena. Se puede aplicar la técnica de Sobel sobre la imagen directamente (Fig. 7a), aunque se recogen mejores resultados si antes se suaviza con algún método de filtrado como el gaussiano (Fig. 7b). Incluso después de la simplificación obtenida tras este procesado, si la imagen final, que será entrada de nuestro prototipo de optófono, es suficientemente grande, se reducirá su tamaño a uno más práctico, que guarde un compromiso entre su definición y el sonido resultado de su sonificación. Este tamaño suele rondar 32x32 o 64x64 píxeles para imágenes cuadradas, aunque no está limitado a este tipo (Fig. 7c, 7d).

La heurística adoptada es la de suavizar en primer lugar la imagen inicial y realizar la detección de bordes mediante el método ya descrito. La imagen de bordes resultante (Fig. 8a) se puede binarizar mediante un umbral de forma que obtengamos finalmente dos únicos niveles de intensidad posibles para cada píxel: blanco o negro (Fig. 8b). Sin embargo, mediante esta práctica evitamos trabajar con múltiples niveles de gris perdiendo uno de los dos grados de libertad de nuestro sonido estructurado, el volumen, dado por la amplitud de las ondas

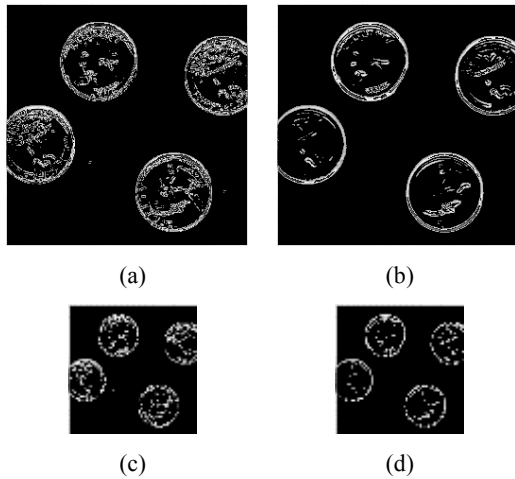


Fig. 7: a) Detección de bordes sin suavizado previo; b) detección de bordes con un suavizado gaussiano previo; c) y d) imágenes a) y b) respectivamente, redimensionadas.

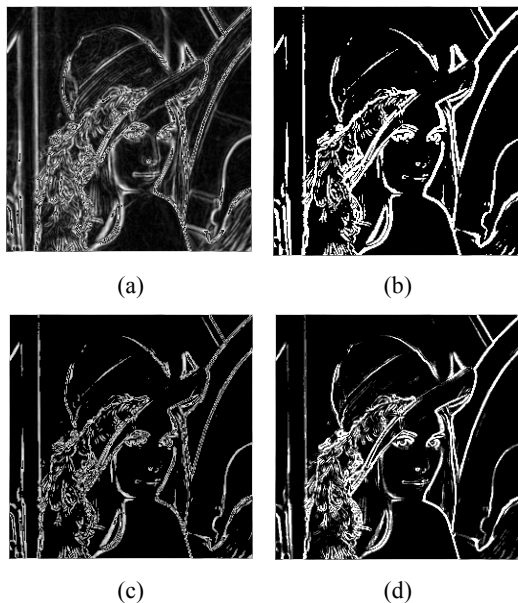


Fig. 8: a) Imagen de bordes; b) imagen de bordes binarizada; c) imagen de bordes umbralizada; d) imagen de bordes umbralizada, ecualizada y cuantizada

generadas. Otra posibilidad más integradora es utilizar el umbral tan solo para transformar a negro aquellos píxeles cuyos niveles sean inferiores a dicho valor y dejar intactos los superiores, lo que llamamos detección de bordes umbralizada en lugar de binarizada (Fig. 8c). En este caso la imagen terminará con múltiples y diferentes niveles de intensidad y, en general, tendrá la posibilidad de crear diversas amplitudes de onda. Esta versión de procesamiento se puede mejorar aún más ecualizando el histograma, que repartirá los niveles de intensidad supervivientes en nuestro intervalo de valores permitidos y ampliará las diferencias en los niveles de gris, por tanto también las amplitudes de las ondas. Por último, si además de esta ecualización, contamos con una reducción de los niveles

de gris resultantes, obtendremos lo que llamamos imagen de bordes umbralizada, ecualizada y cuantizada (Fig. 8d), siendo la técnica que mantiene mejor el compromiso entre las dos características sonoras y la dificultad de interpretación de la sonificación.

Para comprobar la bondad de la sonificación se han creado baterías de pruebas con imágenes sintéticas y se ha experimentado con 28 voluntarios sin discapacidad visual. El porcentaje de personas que consigue identificar más de un 50% de las imágenes presentadas en forma sonora llega hasta el 71% de los casos y un 21% de las personas entrevistadas identificaron más del 75% de los objetos sonificados. Teniendo en cuenta que los voluntarios no habían sido entrenados previamente consideramos que estos resultados son altamente satisfactorios. Sin embargo, nuestra hipótesis respecto al uso de las escalas musicales natural y cromática no ha podido ser contrastada estadísticamente hasta el momento de la escritura de este trabajo.

## 5. CONCLUSIONES Y FUTUROS TRABAJOS

Este trabajo presenta un prototipo de optófono que es capaz de sonificar imágenes digitales en niveles de gris usando los niveles de intensidad y la distribución espacial de sus píxeles. La mayor dificultad en la sonificación de imágenes es el exceso de información que contienen para describir una escena, fundamentalmente dado por el número y distribución de los píxeles. Por esta razón aplicamos conocidos métodos de segmentación y simplificación de imágenes. Los resultados preliminares son satisfactorios y las pruebas realizadas sobre diferentes individuos indican que la asimilación de información sonora, aun no estando exenta de dificultad, mantiene un grado de precisión acorde con las expectativas creadas por adelantado, pues el oyente, sin llegar a extraer finos detalles de la imagen consigue formarse una grosera idea de dicha escena.

No obstante, y aunque el presente trabajo sólo tiene en cuenta el volumen y el tono como características de sonido, es un paso preliminar a la incorporación de una nueva variable adicional, el timbre del instrumento. El timbre es una característica musical determinante e identificativa del sonido de un instrumento y es consecuencia directa de la forma, los materiales y el elemento que origina el sonido. Dado un sonido con una frecuencia principal, la distribución de intensidad de sus armónicos y parciales generan un timbre u otro [8]. Esta característica la podríamos utilizar para sonificar una cualidad inherente y añadida de las imágenes, el color. De esta manera introduciríamos en paralelo la información no sólo de la distribución espacial de los píxeles de la imagen sino también un contenido global a partir de su homogeneidad de color. Para la consecución de este objetivo podríamos hacer uso de un sistema de inferencia borroso pues el color o la tonalidad como característica global de una imagen tiene un alto grado de indeterminación y utilizar una librería MIDI para crear los sonidos de los diferentes instrumentos.

Por supuesto, otro punto clave que queda como futuro trabajo es experimentar este método con personas con discapacidades visuales, entrando en contacto con las instituciones pertinentes.

Por último, cabría destacar el futuro uso del optófono como sonificador de secuencias de imágenes, asumiendo que un excesivo estrés para el oyente mermaría las posibilidades prácticas de aplicación y conllevaría, necesariamente, a emplear

una reducción aún mayor de las características generadoras del sonido.

### AGRADECIMIENTOS

Los autores agradecen al profesor D. Juan José Pantrigo Fernández por los consejos musicales aportados.

### REFERENCIAS

- [1] Blauert, J., *Spatial Hearing: The Psychophysics of Human Sound Location*. MIT Press, 1997.
- [2] Flowers, J.H., Buhman, D.C., Turnage, K. D. "Cross-modal Equivalence of Visual and Auditory Scatterplots for Exploring Bivariate Data Samples", *Human Factors*, 39, 1996, pp. 341-351.
- [3] Gonzalez, R. C., Woods, R. E., *Digital Image Processing*, Second Edition, Prentice-Hall, 2002
- [4] Gulick, W. L., Gescheider, G. A., Frisina, R. D., *Hearing: Physiological Acoustics, Neural Coding, and Psychoacoustics*, Oxford University Press, 1989.
- [5] Hartmann, W.M., *Sounds, Signals, and Sensation: Modern Acoustics and Signal Processing*. Springer Verlag; New York, 1997
- [6] Kramer, G. *Some Organizing Principles for Representing Data with Sound*. Proceeding of the First International Conference on Auditory Display (ICAD), 185-221, 1994.
- [7] Kramer, G. *Auditory Display: Sonification, audification, and auditory interfaces*. Proceeding of the First International Conference on Auditory Display (ICAD), 1994.
- [8] López, M. R., *Ingeniería Acústica*, Ed. Paraninfo, 2000
- [9] Madhyastha T. M., Reed, D. A., *Data Sonification: Do You See What I Hear?*, IEEE Software, Vol 12(2), 85-90, 1995.
- [10] Meijer, P.B.L, *An Experimental System for Auditory Image Representations*, IEEE Transactions on Biomedical Engineering, Vol 39(2), 112-121, 1992.
- [11] Moore, B. C. J., *Handbook of Perception and Cognition*, Vol 6, Hearing, Academic Press, 1995.
- [12] Moore, B. C. J., *An Introduction to the Psychology of Hearing*, 4th ed., Academic Press, 1997.