# A Hierarchical Visual Feature-Based Approach For Image Sonification

Ohini Kafui Toffa and Max Mignotte

***Abstract*—This paper presents a new image sonification system that strives to help visually impaired users access visual information *via* an audio (easily decodable) signal that is generated in real time when the users explore the image on a touch screen or with a pointer. The sonified signal, which is generated for each position within the image, tries to capture the most useful and discriminant local information about the image content at different levels of abstraction, ranging from low-level (at the pixel level) to high-level (segmentation) and combining low-level (color edges and texture), mid-level and high-level (gradient or color distribution for each region of the image) features. The proposed system mainly uses musical notes at several octaves, the notion of timbre, and loudness but also uses pitch, rhythm and the distortion effect in an intuitive way to sonify the image content both locally and globally. To this end, we use perceptually meaningful mappings, in which the properties of an image are directly reflected in the audio domain, in a very predictable way. The listener can then draw simple and reliable conclusions about the image by quickly decoding the sonified result.**

***Index Terms*—Sonification, visually impaired, sound synthesis, auditory feedback, audio mapping.**

## NOMENCLATURE

*Variables*

| | |
|---|---|
| $R_{max}$ | Maximum number of regions |
| $N_{Hue}$ | Bin number of the hue |
| $N_{Grad}$ | Bin Number of the gradient |
| $N_{Samp}$ | Number of samples 16384 |
| $R[][]$ | Regions of Segmented Image |
| $y$ | Current Region label |
| $y_{Old}$ | Old Region Label |
| $Pos_x, Pos_y$ | Cursor Position |
| $H_R[][]$ | Mixture of Hue of size $R_{max} * N_{Hue}$ |
| $G_R[][]$ | Mixture of Gradient of size $R_{max} * NGrad$ |
| $S_R[]$ | Saturation table of length $R_{max}$ |
| $L_R[]$ | Luminance table of length $R_{max}$ |
| $GM_R[]$ | Gradient Mean table of length $R_{max}$ |
| $OG_R[]$ | Oriented gradient table of length $R_{max}$ |
| $N_{Freq}[]$ | Notes Frequencies of Octave $C_4$ of length 7 |
| $M_{Freq}[], P_{Freq}[]$ | Sound samples of length $N_{Samp}$ |

Manuscript received January 10, 2019; revised September 10, 2019 and November 27, 2019; accepted March 30, 2020. Date of publication April 17, 2020; date of current version January 29, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chi-Chun Lee. *(Corresponding author: Ohini Toffa.)*

The authors are with the Vision Laboratory of the Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal, Faculté des Arts et des Sciences, Montréal H3C 3J7, QC, Canada (e-mail: ohini.kafui.toffa@umontreal.ca; mignotte@iro.umontreal.ca).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2987710

## I. INTRODUCTION

SONIFICATION is the translation of data into sound. More generally and precisely, it is the use of non-speech audio to convey information or perceptualize data. In fact, this field has greatly progressed over the past century and currently now constitutes an established area of research. One of the earliest and most successful applications of sonification is the Geiger counter, which was invented in 1908 and which uses the rate of clicking to convey the level of radiation being detected in the immediate vicinity of the device. One of the most recent and technologically advanced applications is SONAR [1], which uses echo location, very similar to that used by bats and marine mammals (whales, dolphin, etc.), to convey information about the 3D underwater environment, not only about the geometry (*i.e.*, position, shape, orientation of one or several near or far objects) but also, to a certain extent, about the surface properties of the detected objects, the nature of the sediments lying on the seafloor and/or the structure of the seabed. With the evolution of the computing technology and the presence of tactile screens, smartphones, tablets and wearable devices, sonification has become more interactive [2] and is used in emergency services, aircraft cockpits, assistive technologies, climate sciences [3], elite sports [4], multimodal interactive environments [5], engineering analyses and simulations and interpretations based on the sonification of physical quantities [6].

Hence, the idea behind image sonification is to find ways to translate the image data, which describe shape, color and texture (sometimes depth information), into sounds. This is a recent field that has naturally emerged after the development of image and sound processing techniques. Such sonification may be particularly useful for finding ways to represent information that would be accessible to users with visual impairments. This technique is also beneficial in circumstances where visual representations would be impossible to use or to enrich a graphical realization [7], for human-computer interactions or for medical applications [8]. In these latter application cases, auditory feedback can complement visual data without requiring a surgeon to constantly monitor the screen or to help him or her to understand critical and additional useful information.

Previous work on image sonification can be roughly divided into two categories. In high-level (symbolic) sonification, visual

information is translated into natural speech language. This field is still in its infancy since it is very difficult today, if not impossible, to fully understand the semantic content of all images. Let us note that the obvious limitation of such sonification is that it is limited to images composed of objects that have obvious semantic representations. For example, it is not clear how to sonify complex shapes, color, textures or variations of these visual cues and abstract drawings and paintings. In contrast, low-level image sonification aims to transpose image features or visual information into an abstract non-verbal audio signal [9]. This work falls into this latter category. Let us also add that this type of sonification can be quite complementary to a high-level sonification type for the previously mentioned reasons. Hence, image sonification can be viewed as a data conversion or a data mapping between the visual and audio domains. Nevertheless, it is crucial to understand that the time-independent two-dimensional nature of an image and the temporal nature of a sound makes this conversion nontrivial, especially since a well-designed sonification system must make intuitive sense, and the listener must be able to effectively extract and discern and real-time interpret important audio features.

In biomedical applications, low-level sonification has often been used for providing audio feedback for heart rate variability, Doppler ultrasound and electroencephalography signals [10] and sometimes used for manual positioning of surgical instruments and surgical navigational system [11]. Few works have been dedicated to image sonification except [8], where a sonification of each segmented nucleus (parameterized by two discriminant statistical geometric feature-based parameters) is presented to the cytologist, in a complementary auditory form, to improve its diagnostic accuracy. Following the same principle, Ahmad *et al.* in [12] use a sonification of optical coherence tomography data, showing images of human breast adipose and tumor tissue, with the aim of distinguishing these tissue types based on the rendered audio signals.

Recent research efforts have been devoted to using low-level image sonification to produce a navigational system to assist and improve blind and visually impaired people's mobility in terms of safety and speed. A portable wearable headgear-based device with cameras located slightly above the position of the eye in [13] and a mobile tablet with two cameras in [14] have been used to build a vision assistance system that uses depth inference *via* real-time stereo matching and a depth-to-sound mapping to inform the user of the surroundings *via* sound. In the vOICe [15], the image is captured using a single video camera mounted on headgear, and the captured image, possibly a depth image as proposed in [16], is scanned from left to right for sound generation (with the sound loudness depending on the brightness of the pixel).

Exploiting disparity data with a sonification system is interesting to assist the mobility of visually impaired people to bypass obstacles and hazards when this depth information is available, and some efforts have been made in this regard, as mentioned above. Nevertheless, few works have been proposed to aid a visually impaired user to recognize objects in a (synthetic or) natural image (or painting) or, more modestly, to help visually impaired persons perceive some characteristics of the main shapes, shown in an image, in terms of color, luminosity and texture, which would allow them then to draw some conclusions about the content of the image (or video frames).

In this context, Martins *et al.* [17] was the first to sonify a single texture pattern with a periodic audio signal. Then, chronologically, Yeo and Berger [18] used an image sonification technique based on simple raster scanning of the image to generate a sound whose loudness linearly fits the brightness value of the scanned grayscale image pixels. A similar approach was used in [19], in which pixel values are translated into a musical notes. Ivan and Radek [20] presented a simple sonification method for mapping color information to a frequency oscillator, where color information was mapped to the wave envelope, waveform and frequency of a sound. Yoshida *et al.* in [9], presented a sonification methodology based on edge gradients and distance-to-edge maps extracted from an image and *via* a mobile touch-screen device. More recently, in [21]–[23], the authors used the concept of color, color mixture with the combination of acoustical entities and the grade of roughness on pre-classified natural regions and edges with drum rhythms in their sonification system. A sonification tool proposed in [24] starts by scanning the loaded image from top to bottom and produces a sound for each row of the image that plays in sequence and that consists of other elementary sounds; more precisely, the authors used the HSV color space, and in this space, the loudness is determined by the luminance or value (V) of the pixel, the signal'spectral envelope is controlled by the saturation (S) (from a sinusoid to a square wave for the lowest to the highest saturation value), and the hue (H) is mapped into the fundamental frequency of the synthesized sound. In [25], different sonification strategies for a guidance task were used to help participants to quickly find a vertical hidden target randomly placed on a virtual horizontal line on a pen tablet. Finally, in [26], the author proposed a mobile application that sonifies HSV color and greyscale images and permits blind children to recognize on an interactive screen a straight line and a curve.

In this work, we present a new image sonification system that strives to help visually impaired users access visual information *via* an audio signal. The visually impaired user can explore the image on a touch screen and receives real-time auditory feedback about the image content at the current position. The proposed system works in real time and is intuitive. We use perceptually meaningful mappings, in which the properties of an image are directly reflected to the audio domain in a very predictable way and can be easily extracted by the listener that can draw conclusions about the image by decoding the sonified result. To this end, the audio signal that is generated tries to capture the most useful information of an image, such as low-level image processing cues (*i.e.*, color edges and texture) and mid-level cues (histogram of the gradient) at a low-level (at the pixel level) and high-level of information obtained from an efficient segmentation algorithm that represents the image content in different sub-parts or regions of coherent textural properties. The proposed method uses mainly musical notes at several octaves, the notion of timbre, and loudness but also the pitch, rhythm and distortion effect in an intuitive way to sonify locally and globally the image content. Such system could be useful to visually

impaired persons by providing a special translation experience of paintings in a digital museum or an interpretation of the image of a live event received on a mobile phone.

## II. Sonification Model

A well-designed real-time sonification system must be fast since it computes a sound for a position that varies in the image as a *probe* while presenting the result by means of either headphones or speakers. The sonified audio result must capture as much reliable information as possible about the image on many levels of abstraction, ranging from low-level to high-level and combining low-, mid- and high-level features at each location of the image. Above all, the sonified sound must be logical and consistent and make intuitive sense. The listener must be able to effectively and quickly extract (*i.e.*, interpret in real time) important and discriminant visual features of the image, easily and intuitively identifiable, from the sonified result in order to draw simple and reliable conclusions about the image content. To do that, we must propose perceptually meaningful mappings in which the properties of an image are directly reflected in the audio domain in a very predictable way.

Our sonification model first relies on segmentation of the image. This allows us to obtain a high-level representation of the image to be sonified in which different sub-parts or homogeneous regions of coherent textural properties are located. To this end, we suggest using the reliable segmentation model proposed in [27], which is freely available on the Web (although any other model could be used). This method is based on the combination, in the Variation of Information (VoI) sense, of quickly and roughly estimated segmentation results obtained by the simple $K$-means procedure when the image is expressed in different and complementary color spaces. For each identified region or subpart of the image, a different audio sound, lasting one second, and repeating itself in a loop as long as the *pointer* (whose position is controlled by the mouse, the keyboard or a touchscreen device) remains in this region, will be generated with different characteristics, which we now explain.

### A. Choice of the Color Space

First, we have to select a suitable color space in which we intend to extract our low- and mid-level discriminating visual features on each presegmented region of the image. In this sonification system, the HSL color model is used as an ideal intuitive color model, which better describes the human perception of color than the RGB model [28]. It was designed in the 1970s by computer graphics researchers to more closely align with the manner in which human vision perceives color-making attributes. In fact, the HSL color space can be easily understood, since it is the color space that can best be explained with words or simple concepts. Moreover, this is confirmed by the fact that this color space is also used by painting or drawing artists to naturally describe a color or to describe the manner in which paints of different colors mix together. In this HSL color space (see Fig. 1), the visual properties of a color can be described with words or simple concepts, such as Hue, Saturation and Luminance. Note that any color space based on the Munsell color system [28], like HSV, is a good candidate since it provides almost the same visual properties.

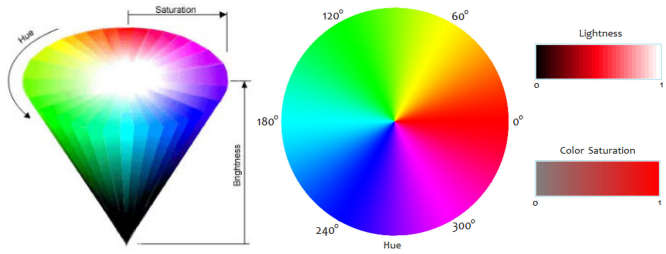


Fig. 1. The HSL color space with the hue component that is arranged in a radial slice (starting at the red hue at $0^o$, passing through green at $120^o$ and blue at $240^o$, and then wrapping back to red at $360^o$) around a central axis of neutral colors, which ranges from black at the bottom to white at the top. The HSL representation models how colors mix together, with the saturation dimension resembling various shades of brightly colored paint and the luminance (or brightness) dimension resembling the mixtures of those paints with varying mounts of black or white paint.

*1) Hue:* is a term that one visually thinks of as an existing color that can be described with simple words, such as *'red'*, *'yellow'*, *'green'* or *'purple'*. Red (or green) is a distinct pure, primary hue, while the hue '*yellow*' is composed of equal quantities of (primary hues) red and green.

*2) Saturation:* is a measure of how intense or pale a color appears, and this concept is governed by the amount of white it contains. This term is generally used to describe the purity of a color. For example, *'pink'* is a tint of the color *'red'* to which between $10\%$ and $70\%$ of white has been added.

*3) Luminance:* Luminance can also be described by words such as bright or dark. This concept is dependent on the amount of energy that is being radiated. Thus, darkness is a lower-intensity shade of a bright red.

This color space is suitable for describing a gradient or color variation in space or time. For example, it can be seen that a cookie becomes more brown as it is baked: its hue remains relatively constant, but its luminance and saturation change during cooking [29].

### B. Choice of the Frequency Sampling

In our application, we use a sampling frequency of $f_e = 16384$ samples/second, which allows us to model a maximum, the Nyquist frequency of 8 kHz, that is used in most modern VoIP (Voice over Internet Protocol) communication products and which is sufficient to model complex audio signals. In addition, human sensitivity to frequency information above 8 kHz is rather limited, and conveying information above this high frequency remains perceptible by healthy human ears but is very sensitive to environmental external noise and thus difficult to decode [30].

Since we want to generate an audio signal that lasts $T = 1$ second with $f_e$, we have to generate a total of 16384 sound samples, which is also a power of two ($16384 = 2^{14}$) and which will allow us to then efficiently use an inverse FFT Fast Fourier Transform (since our sound mapping will be generated in the frequency domain) and to fully give 8 octaves with the highest

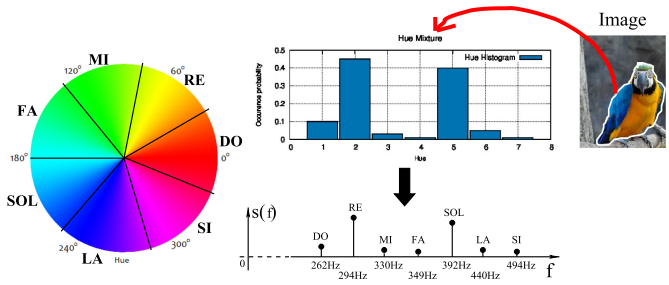


Fig. 2. Hue Translation: the re-quantized seven-bin hue histogram of each region is first estimated and then converted into an impulse function weighted by the corresponding value of the histogram, in the frequency domain, centered on the frequency corresponding to the different notes of the musical game.

frequency given by a 108-key grand piano. With this specification, let us note that the frequency resolution of our sonification model is $\Delta f = 1$Hz.

### C. Sonification Mapping

The different steps of our sonification approach are the following:

*1) Hue Translation:* The core of our sonification mapping is based on the seven musical notes (*do-re-mi-fa-sol-la-si*) possibly played on several octaves at once (this will be explicit in Section II-C3). This choice comes from the fact that several blind or visually impaired persons naturally develop a *musical ear* (certainly due to the cortical plasticity [31][1]) and are able to easily identify every note immediately and in isolation from other played notes [32], [33].[2]

Importantly, HSL offers the possibility to code a color using a precise note with only the H color channel (hue) and linearly (more precisely radially around the circle of the cone) with a semantic (well-understood) expression (*i.e.*, 0-Red, 60-Yellow, 120-Green, 180-Cyan, 240-Blue, 300-Magenta, 360-Red) (see Fig. 2). Additionaly, in this color space, colors with the same hue can then be distinguished semantically with adjectives referring to their saturation and lightness, just as our sonification system will be able to do but with different sound characteristics (such as octave and harmonics), as will be explained later.

More precisely, in our application, the hue information of each region is first modeled by a normalized re-quantized histogram with seven (one for each note) equal-width bin (in the hue interval) as a hue feature vector. In this simpler model, the texture of each pre-segmented region is herein characterized by a mixture of hues, or more precisely, by the values of the re-quantized hue histogram. This model is simple, quick to compute, and allows significant data reduction while being robust to noise and local image transformations.

[1]A part of the core area of the auditory cortex was found to be enlarged by a factor of 1.8 in the blind compared with sighted humans. Such cortical reorganization may be a consequence of the absence of visual input in combination with enhanced auditory activity.

[2]The authors show that blind people perform better than sighted individuals at tasks related to pitch discrimination and pitch-timbre categorization and on a range of auditory perception tasks. This advantage was observed only for individuals who became blind early in life.

Once the seven-bin histogram is computed, each of these seven re-quantized histogram values is converted into an impulse (or Dirac delta) function weighted by the corresponding value of the re-quantized histogram (see Fig. 2) in the frequency domain. More precisely, the first, second, . . ., seventh values of the histogram are converted to an impulse function centered on the frequency corresponding to the different notes of the musical game, *i.e.*, 262Hz (Do), 294Hz (Re), 330Hz (Mi), 349Hz (Fa), 392Hz (Sol), 440Hz (La), and 494Hz (Si), respectively (represented by the white keys of a piano keyboard for the octave $C_4$ Middle C [34]), in a manner such that in the temporal domain, the mixture of hue of each region will be translated by a mixture of pure tones or musical notes (easily identifiable by a visually impaired person) played together in a manner that sounds harmonious and determining the the so-called *pitch* of the generated sound.

*2) Luminance Translation:* Since this concept is dependent on the amount of energy that is being radiated (cf. Section II-A3), we thus bring this concept closer to the different octaves of vibration of a piano. To this end, the luminance value, initially in the interval $[0-255]$ is divided into 8 equal intervals, and each interval is assigned the name of an octave scale $C_n$ [34], in which is played the musical notes defined in Subsection II-C1 with:

$$C_n = \left\lceil \frac{l_R}{32} \right\rceil \tag{1}$$

where $l_R$ is the mean luminance value of the region and $\lceil . \rceil$ the ceil function. For example, if $l_R = 100$, it implies the octave $C_4$ with the standard so-called *Middle C octave* [34] comprised within the range $[\text{Do} = 262\text{Hz} - \text{Si} = 494\text{Hz}]$. If $l_R = 180$, it implies the octave $C_6$ with the so-called *Soprano C octave* using the range $[\text{Do} = 1046\text{Hz} - \text{Si} = 1976\text{Hz}]$.

*3) Saturation Translation:* Since this concept describes the purity of the color (high-saturation colors look rich whereas full- and low- saturation colors look dull and grayish; see Fig. 1), we translate this concept in audio space by adding to the sound previously generated more (for low-saturation colors) or fewer (for high-saturation colors) harmonics, thus making the sound more or less pure. More precisely, the saturation value, initially in the interval $[0-255]$, is divided into 8 equal intervals, and $N_s$ is computed as follows:

$$N_s = 7 - \left\lfloor \frac{s_R}{32} \right\rfloor \tag{2}$$

where $s_R$ is the mean saturation value of the region and $\lfloor . \rfloor$ the floor function. $N_s$ represents the number of octaves; in addition to the one in that is playing the sound previously generated (see Subsection II-C2), we duplicate this mixture of notes on several other (closest) octaves. For example, if $s_R > 224$ implying $N_s = 0$, the pure sound created on only one octave is generated. If $s_R = 100$ implying $N_s = 4$; 4 supplementary octaves (the closest to that estimated in Subsection II-C2) are added to the initial sound with a weighting amplitude of $1/N_s$ (*i.e.*, $1/4$ for our example (see also the example given in Fig. 3), making the generated sound less pure.

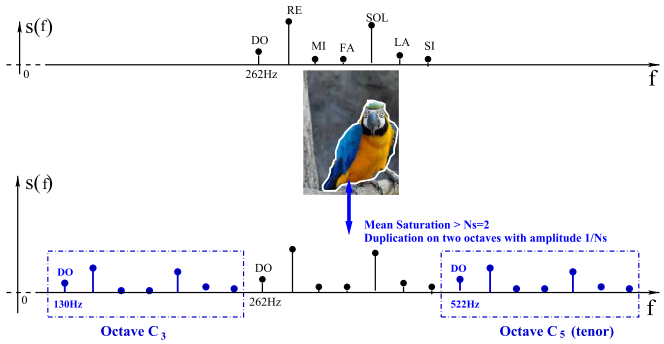


Fig. 3. Saturation Translation: In addition to the mixture of musical notes generated by the hue value (see Fig. 2), we duplicate it on several other (closest) octaves according to the saturation value. For example, if $s_R = 170$ implying $N_s = 2$ (see Eq. (2)), we duplicate this mixture on the two closest octaves of $C_4$ (with a weighting of 0.5).

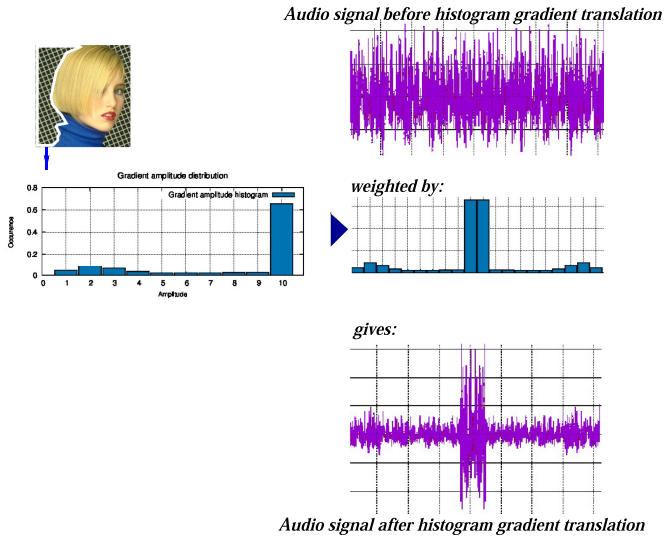


Fig. 4. The histogram or distribution of the amplitude of the gradient module within the considered region (to be sonified) is used to weight the temporal envelope of the audio signal.

*4) Roughness Texture Translation:* In order to aid the user's understanding of the possible roughness (due to the presence of gradients) of a textured region to be sonified, we can alter the purity of the sound signal, thanks to the concept of distortion. This can produce a *vibrant*, *rhythmic*, *growling*, or *gritty* tone depending on the type of distortion used. This effect can efficiently (and intuitively) model the grade and style of roughness of each pre-detected region.

More precisely, we can consider two different types of roughness properties of a region that can be quantified by two statistical features. The first one is the mean of the module of the first-order gradient within the region (defining the global roughness of the region), and the second is the variance of the orientation (following the four main directions) of the (first-order) gradient module (hence quantifying the presence of man-made geometric structure, such as a manufactured object or an un-natural texture).

- In our application, the first type is created by magnitude distortion by adding (noisy) randomized harmonics. This can be simply done by adding, to the frequency bins with null amplitude, randomized values within the interval $[0\ ,\ \rho]$ with $\rho$ proportional to the mean of the module of the first-order gradient within the region.
- The second type can be created by phase distortion by adding (noisy) randomized value $[-\beta\ ,\ \beta]$ to the phase spectrum, with $\beta$ proportional to the variance of the oriented (in the four directions) module of the first-order gradient within the region.

*5) Translation Into Temporal Space:* Once the hue, saturation and luminance and the mean gradient module of each region have been mapped in the frequency domain and expressed in in terms of the magnitude spectrum vector[3] and the variance of the oriented gradient in the phase spectrum vector (thus defining the audio signal's spectral envelope (which is related to the perception of *timbre*), after Hermitian symmetry is imposed on the magnitude and on the phase spectrum, we return to the time domain, with a simple inverse fast Fourier transform (FFT), to get $s(t)$.

*6) Histogram Gradient Translation:* In Section II-C4, we have sonified, *via* the distortion effect of the audio signal, the mean and the orientation variance of the gradient magnitude as two different features related to the gradient-based region activity. Another important visual cue that remains to be expressed by sonification is the histogram or distribution of the amplitude of the gradient module. A way to express this visual cue and to give it a meaningful, interesting audio effect is through the notions of rhythm and loudness of the sonified sound. To this end, and in order to sonify this information to the user, such that it is easily decodable, we use the following strategy: we first compute a re-quantized histogram using ten equal-width bins of the first-order gradient module and use this histogram followed by its mirror projection to weight the 1-second temporal envelope of the audio signal $s(t)$ (see Fig. 4).

More precisely, in our application, we keep a percentage $p$ of the original signal (characterizing and encoding the image low-level visual features listed in previous points 1 to 5), and the weighting is applied for the other $100\% - p$ of the signal. This allows us to avoid generating signals of total silence, as in the example of the region given in Fig. 4. See Algorithm 1 for implementation details.

## III. Experimental Results

In our experiments, we have tested our sonification algorithm on some images from the Berkeley segmentation database (BSD300) [35]. This image-base has both a great variability of naturally colored and textured images and a good (manually hand-segmented) segmentation for each image. This allows us to objectively analyze, discuss and highlight the pros and cons of just our sonification process. Nevertheless, in the absence of a

[3]Algorithmically, since $\Delta f = 1$Hz, in our application (with a sampling frequency of $f_e = 16384$ and 16384 sound samples; see Section II-B), it boils down to filling a 1D vector of length 16384 by simply putting an amplitude value in the n-th cell for the frequency nHz.

**Algorithm 1:** Image Sonification

**Initialization**
Load Image and Convert to HSL
Compute or Load Image Segmentation to $R$
**for** *each region* $r < R_{max}$ **do**
- $H_R[r] \leftarrow ComputeHueHistogram()$
- $G_R[r] \leftarrow ComputeGradientHistogram()$
- $S_R[r] \leftarrow ComputeSaturation()$
- $L_R[r] \leftarrow ComputeLuminance()$
- $GM_R[r] \leftarrow ComputeGradientMean()$
- $OG_R[r] \leftarrow ComputeOrientedGradient()$

**end**
**Wait For Event**
**while** *user input and not exit* **do**
  $Pos_x, Pos_y \leftarrow GetCursorPosition()$
  $y \leftarrow R[Pos_x][Pos_y]$
  **if** $y <> y_{Old}$ **then**
    *GenerateSound()*
    **for** $i < N_{Hue}$ **do**
      1. **Translate hue and luminance**
      $C_n \leftarrow \lceil \frac{L_R[y]}{32} \rceil$
      $M_{Freq}\lfloor N_{Freq}[i] * 2^{C_n-4} \rfloor \leftarrow H_R[y][i]$
      2. **Translate saturation**
      $N_s \leftarrow 7 - \lfloor \frac{S_R[y]}{32} \rfloor$
      **for** $l < N_s$ **do**
        $M_{Freq}\lfloor N_{Freq}[i] * 2^{Neighbors(C_n,l)-4} \rfloor \leftarrow H_R[y][i]/N_s$
      **end**
    **end**
  **for** $k < N_{Samp}/2$ **do**
    3. **Magnitude Distortion**
    **if** $M_{Freq}[k] = 0$ **then**
      $\alpha = 5 * rand() * GM_R[y]$
      $M_{Freq}[k] \leftarrow \alpha / N_{Samp}$
    **end**
    4. **Phase Distortion**
    $\beta = 5 * rand() * OG_R[y]$ $P_{Freq}[k] \leftarrow \beta / N_{Samp}$
  **end**
    $HermitianSymmetry(M_{Freq}, P_{Freq})$
    $Sound \leftarrow IFFT(M_{Freq}, P_{Freq})$
    5. **Translate Gradient Histogram**
    $Sound \leftarrow Weight(Sound, G_R)$
    $PlaySound()$
    $y_{Old} \leftarrow y$
  **else**
    *PlaySoundIfNotPlaying()*
  **end**
**end**

segmentation map for each image, we can use any automatic segmentation model and especially the one proposed in [27] which obtains a segmentation score, in terms of the Rand Index equals to 0.81, meaning that on average, $81\%$ of pairs of pixel labels are correctly classified compared to the segmentation maps of the BSD300, considered as ground-truths.

In our tests, we set $\rho$ and $\beta$ (Section II-C4), 5 times the mean of the module of the first order gradient and 5 times the variance of the oriented module of the first-order gradient, respectively. A high value of $p$ (Section II-C6) reduces the effect of the gradient interpretation on the signal envelop, while a small one increases the risk of signal with silence. We then used $p = 10\%$ as a good trade off. Note that those values are empirical.

### A. Discussion

When we position the pointer (controlled by the mouse or the keyboard) in the middle bottom and left border of the image shown in Fig. 5a, we can easily recognize and thus localize two regions associated with a pure tone sound (lasting one second, and repeating itself in a loop) with a specific frequency corresponding to the musical note *Do* for the red part of the wool turtleneck sweater of the woman and the specific note *La* for its two blue parts (at the left arm and neck) (see Section II-C1). We can easily also guess the homogeneous black part of the sweater since the emitted sonified sound mainly vibrates at the bass tones (see Sect. II-C2 (low frequencies), but with several other harmonics (with smaller amplitudes) (cf. Sect. II-C3), indicating the presence of a very low luminance and very saturated colors such as black. Let us note that this part can be distinguished from the background that is dark since the latter region is a uniform dark region without gradient (unlike the sweater region), and thus the temporal envelope of the sonified sound is different (see Fig. 5b). We can also easily localize the blond hair of this person since the sonified sound is also a pure tone corresponding to the *Re* musical note, but with a slight magnitude distortion effect due to the presence of a mean gradient in this particular region (cf. Sect. II-C4) and thus sonifying the particular textural roughness of this region and also distinguishing the hair region from the facial area. Indeed, the lightness of this part is more important, the saturation is lower, and the distribution of the gradient is radically different (see Fig. 5b), thus generating a very different temporal envelope for the sonified signal. We can easily draw the outline of the person without ambiguity. Finally, the background, whose area looks like a sort of fence, is associated with a sound that is a very complex sound (highly saturated) with a lot of (amplitude and phase) saturation (cf. Sect. II-C4) and with a very peculiar temporal envelope, creating a kind of scratchy noisy sound that grumbles regularly every half second and thus representing appropriately the regular grilling.

Fig. 6 shows four images from the BSD300 Berkeley database with some audio samples generated at different locations of these images. Figures (a) and (d) share the same semantic concepts as do (b) and (c). We can notice that the audio results generated at the sky of the second and third images are very similar and very characteristic. Similarly, the sonified sound generated for the modern building in the second and third images are very similar (despite a slightly different texture, demonstrating that our proposed sonification system generalizes well across regions from the same semantic concept or label) and also characteristic of a man-made structure with geometric and regular patterns. The sound emitted by the by the starfish or its background is very rich and complex with amplitude and phase distortion

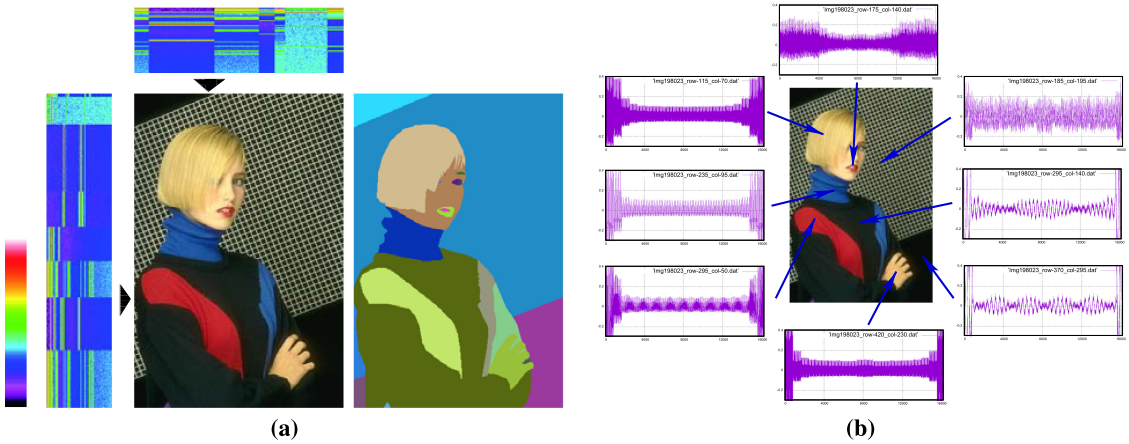


Fig. 5. **(a)**: Image number 198023 from the BSD300 Berkeley database [35] (left) and its segmentation (right) with the spectrogram of the generated sound when the user examines the image from left to right starting from line 320 and from top to bottom starting from column 110. The frequency resolution of the spectrogram is $\Delta f = 20$Hz, and the spectrogram ranges from 0Hz to 2 kHz (horizontally for the left one and vertically for the top one); the data are represented with the thermal (false-) color scale shown on the far left. **(b)**: Some audio samples generated at different locations of the image.

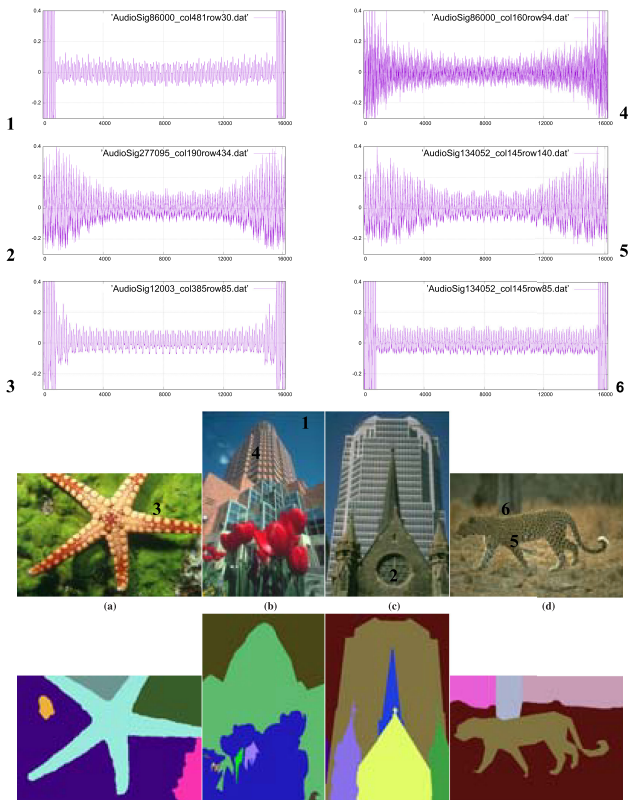


Fig. 6. Images number **(a)** 12003, **(b)** 86000, **(c)** 277095, and **(d)** 134052 from the BSD300 Berkeley database [35] with some audio samples generated at different locations of these images.

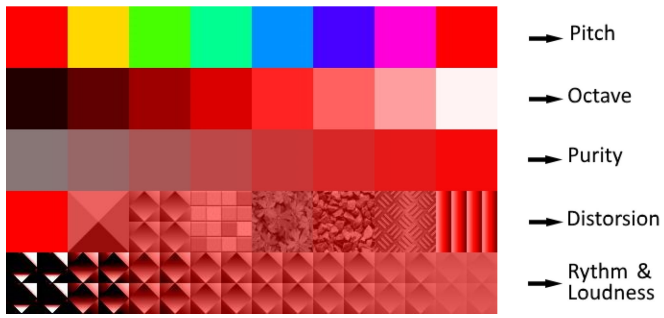


Fig. 7. Experiment I: Calibration and training image.

characteristic of complex textures. Nevertheless, we can hear easily in them the musical notes Do-Re for the starfish and Mi-Fa for the background in relation to their respective hue.

## B. Validation

To validate the sonification model, we obtained a certificate of ethics from the Université de Montréal and performed a pilot study (easily reproducible, from the BSD300, to facilitate eventual further comparisons with future methods) with 14 volunteers (students and nonstudents). Note that each test lasted approximatively 1 h, and each participant was involved in a minimum of two tests. This puts constraints on a personś availability and capacity to stay motivated along all the tests and reduced the number of different participants and the number of test samples. Due to some constraints, we were unable, unfortunately to include blind persons in the study. However, we believe that if sighted people are able to perform well on our system, blind people will do better since they demonstrate better ability with sounds [32], [33]. In all the experiments we masked the images to the user with a black screen.

The study consisted of multiple experiments: mapping properties recognition, scene description, form detection, image categorization and a longitudinal study. In the experiments all subjects explored the very same images but with different orderings to avoid presentation order effects. Only one of the subjects had a *musical ear*, and none of them had a training session beforehand.

*1) Experiment I (Mapping Properties' Recognition):* A calibration image (Fig. 7) containing five rows (pitch, octave, purity, distortion and loudness) was presented to the subject in order to train him with the system. For each row, the model related property was activated, and the subject scanned the row horizontally in order to learn the behavior of the sound based on the property.

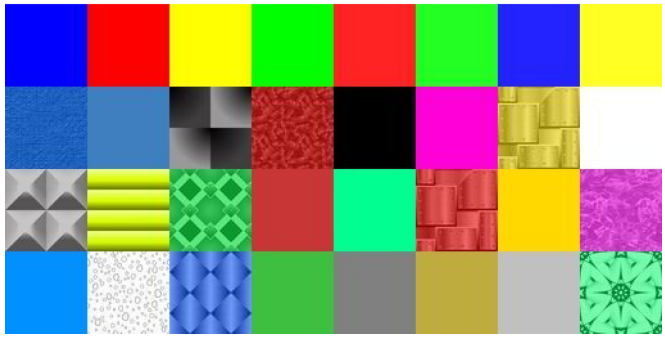

Fig. 8. Experiment I: Mosaic of testing images.

TABLE I
RESULTS OF EXPERIMENT I

| Property | Question | Single | All |
|---|---|---|---|
| Pitch | What is the color of the square ? | 62% | 51% |
| Octave | Is the square dark or light ? | 96% | 75% |
| Purity | Is the square pure or dirty ? | 83% | 66% |
| Texture | Does the square have a texture ? | 92.0% | 72% |
| Loudness | Is the image contrasted ? | 90% | 69% |

The learning session of the selected property lasted approximatively 5mn. Immediately after this session, the user was tested on the selected property using a dozen square images, masked by a black screen, from the calibration (Fig. 7) and mosaic (Fig. 8) images.

After all the properties had been separately tested, we activated all the properties in order to test the effects of the combination. The user was presented each square of the mosaic (Fig. 8) for identification.

Five volunteers participated in the experiment, and the results of the testing are reported in Table I.

As we can observe, the detection of the colors using the pitch, was difficult for the participants while detecting other properties was easy. This is explained by the fact that distinguishing between low and high frequencies or the level of the volume is easier than detecting 7 musical notes. With more training time and practice, a better result could be achieved as we will demonstrate in Section III-B5.

We observed that combining all the properties slightly affected the precision of each property, especially, for a nonmusical ear. For example, a low-saturation image, introduces higher octaves (*via* the added harmonics), which makes the image sounds more acute, as does having a higher luminosity. Thus, it is sometimes difficult to identify whether the acuity of the generated sound is due to the luminosity or saturation of the source image (*i.e.*, the pitch of the musical note or the presence of harmonics). We also observed that for a nonmusical ear, it is difficult to detect the color (or the pitch) when the sound is mainly composed of very low or very high frequencies.

*2) Experiment II (Scene Description):* In experiment II, three participants from experiment I were given 5 minutes to explore the images shown in Fig. 6, without further information. These images were selected to check whether their description fits the interpretation we performed in Section III-A. At the end of the exploration of each image, the participants had to provide an oral description of the scene as they imagined it. A qualitative evaluation is reported in Table II. We can notice that the subjects were able to easily recognize the shapes of the objects, but failed sometimes to recognize the color when the object was highly textured or too dark.

TABLE II
RESULTS OF EXPERIMENT II

| Image | Subject | Description |
|---|---|---|
| (a) | 1 | A yellow textured object with spikes on a dark background. Could be a flower or a homehouse. |
| | 2 | A bright textured form with spikes on a dark background. No idea of what it could be. |
| | 3 | A very textured clear form in the shape of a star on a green background. Green sky? No idea of what it could be. |
| (b) | 1 | A vertical pointed shape with a light red texture on the top and a dark one somewhere on the bottom. A blue background on the top. Could be a pyramid or a head with a pointed hat. |
| | 2 | A vertical textured shape with a blue colored background on top. Could be a bust of a man. |
| | 3 | A light red texture in the shape of a bottle. A clear background with no texture on the top. Could be a bottle or a flower. |
| (c) | 1 | A centered vertical shape textured in red. A clear blue background on the top. Could be a trunk. |
| | 2 | A standing elongated red form, very textured. Could be a tree trunk. |
| | 3 | A vertical dark textured shape. A clear background with no texture on the sides and top. No idea of what it could be. |
| (d) | 1 | A horizontal red textured shape on a green dark background. No idea...a landscape? |
| | 2 | A horizontal bright textured shape centered on a green dark background. Could be a bird in the countryside. |
| | 3 | A light textured shape centered on a dark background. Could be a homehouse. |


Fig. 9. Experiment III: Image numbers **(a)** 66075 (a long ostrich's neck in a plain), **(b)** 101085 (three vertical sculptures in a garden), **(c)** 227092 (a vase laid against a wall), and **(d)** 242078 (six umbrellas on a terrace) from the BSD300 Berkeley database [35].

*3) Experiment III (Form Detection):* During this experiment, eleven subjects had a training of 5 min with the mosaic image (Fig. 8). Only two of them were also involved in experiments I and II. We selected from the database four images (Fig. 9) that contained different objects in terms of shape and number. The participants were given 5 min to explore each image. They had to then say, based on their exploration, if the image contained a long ostrich's neck in a plain, three vertical sculptures in a garden, a vase laid against a wall or six umbrellas on a terrace. The goal of this experiment was to check if the information

TABLE III
CONFUSION MATRIX OF EXPERIMENT II

| | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| (a) | **0.55** | | 0.27 | 0.18 |
| (b) | | **0.91** | | 0.09 |
| (c) | 0.36 | | **0.64** | |
| (d) | 0.09 | 0.09 | 0.09 | **0.73** |

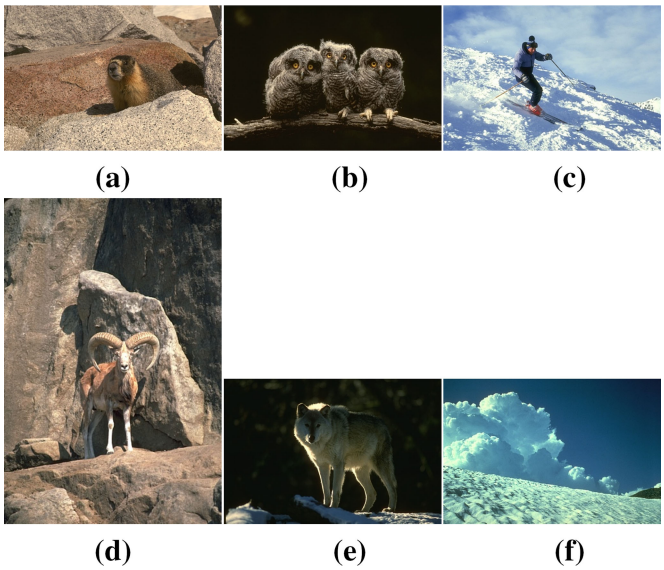


Fig. 10. Experiment IV: Image numbers **(a)** 41069, **(b)** 42044, **(c)** 61060, **(d)** 304074, **(e)** 42078, and **(f)** 176039 from the BSD300 Berkeley database [35] grouped vertically by visual similarity.

about the regions and edge detection was properly conveyed in the model.

The confusion matrix for the results of test III is shown in Table III. Most of the subjects were able to associate the images to the correct content. The greatest confusion was for the vase and the ostrich's neck, since both have a similar vertical shape in the subject's imagination. The content that was easily identified was the three vertical sculptures in a garden (image (b)).

*4) Experiment IV (Image Categorization):* During this experiment, three pairs of images (Fig. 10) that look similar were presented (5 min per image) to the eleven participants from the previous experiment. They had to group them into pairs based on their global similarity in terms of the sounds heard. This third test permitted validating the mapping of the whole hierarchical visual features (regions, edge, colors and texture) into sounds.

This test was more difficult than the previous one because it required the subjects to memorize many tones per image (3 on average). The association between the two images was not straightforward since several tones characterize an image, and the challenge was to use the dominant tones as a reference. The results presented in Table IV indicated that some people associated a bright image with a dark one because they heard one low-frequency sound in some part of the first one. Others associated a blue tinted image with a red one because both contain high-pitch sounds, and it is not easy to distinguish

TABLE IV
ASSOCIATION MATRIX OF EXPERIMENT IV

| | (a) | (b) | (c) | (d) | (e) | (f) |
|---|---|---|---|---|---|---|
| (a) | X | | 0.09 | **0.73** | 0.09 | 0.09 |
| (b) | | X | | 0.09 | **0.73** | 0.18 |
| (c) | | | X | 0.09 | 0.18 | **0.64** |
| (d) | | | | X | | 0.09 |
| (e) | | | | | X | |
| (f) | | | | | | X |

musical notes at such high frequencies. However, the majority of people grouped together the red tinted (a) and (d) images (high-pitched Do sound), the dark (b) and (e) images (very saturated low-frequency sound), and the blue-dominant (c) and (f) images (high-pitched La sound).

*5) Longitudinal Study:* Two participants who were involved in all four previous experiments were allowed to use the system for a long time in order to observe their progression and learning curve. After they had used the system for two or three additional hours, they were able to easily distinguish the colors using the pitch. They improved their result on the combination testing (Section III-B1) by an average of 20%. They appreciated the research in the following terms:

Participant 1: *I improved myself by taking the calibration test multiple times. I liked the mosaic testing, but it was with the scene description experiment that I better understood the usefulness of the research. This could be helpful for visually impaired people.*

Participant 2: *I especially liked the scene exploration tasks (Experiments II to IV). It made me more imaginative and helped me see the utility of the project.*

## IV. CONCLUSION

In this paper, we have presented a new image sonification system that provides an intuitive mode of obtaining visual local spatial information and some context information about an image for visually impaired persons. The proposed system uses a set of hierarchical visual features about the image content at different levels of abstraction and perceptually meaningful mappings based on the additive synthesis technique, in the spectral domain; it uses the concepts of timbre, loudness, pitch, rhythm and different distortion effects to translate the appearance of each individual pre-segmented region of the image into the audio domain. The proposed system allows us to easily localize different regions and classify regions into man-made and natural regions, sometimes with automated man-made object recognition. This system can be complementary to high-level image sonification (*i.e.*, an automatic verbal translation model), which is prone to errors and with which the listener can also miss all the richness, subtleties and complexities of the underlying visual information.

The validation results showed that although the subjects did not have a *musical ear* and did not have any training session, in some cases, they were able to detect objects in the images and group images based on the visual features translated into sounds. This also showed that users were able to improve their

performance on the system with more practice. These results are promising since people with visual impairments (*musical ears*) and some training sessions will surely be able to do better.

## REFERENCES


[1] M. Mignotte, "Unsupervised sonar image segmentation with a hierarchical Markovian approach and classification of cast shadow shapes with statistical models," Ph.D. dissertation, French Naval academy, Brest University, Brest France, Jul. 1998.

[2] N. Degara, A. Hunt, and T. Hermann, "Interactive sonification [guest editors' introduction]," *IEEE MultiMedia*, vol. 22, no. 1, pp. 20–23, Jan. 2015.

[3] V. Goudarzi, "Designing an interactive audio interface for climate science," *IEEE MultiMedia*, vol. 22, no. 1, pp. 41–47, Jan. 2015.

[4] N. Schaffert and K. Mattes, "Interactive sonification in rowing: Acoustic feedback for on-water training," *IEEE MultiMedia*, vol. 22, no. 1, pp. 58–67, Jan. 2015.

[5] A. Tajadura-Jiménez, N. Bianchi-Berthouze, E. Furfaro, and F. Bevilacqua, "Sonification of surface tapping changes behavior, surface perception, and emotion," *IEEE MultiMedia*, vol. 22, no. 1, pp. 48–57, Jan. 2015.

[6] G. Dubus and R. Bresin, "A systematic review of mapping strategies for the sonification of physical quantities," *PLoS One*, vol. 8, no. 12, Dec. 2013, Paper e82491.

[7] K. Franklin and J. Roberts, "Pie chart sonification," in *Proc. IEEE Inf. Visualization (IV03)*, Jul. 2003, pp. 182–196.

[8] A. D. N. Edwards, G. Hines, and A. Hunt, "Segmentation of biological cell images for sonification," in *Proc. Congr. Image Signal Process.*, May 2008, vol. 2, pp. 128–132.

[9] T. Yoshida, K. M. Kitani, H. Koike, S. Belongie, and K. Schlei, "Edgesonic: Image feature sonification for the visually impaired," in *Proc. 2nd Augmented Human Int. Conf.*, 2011, pp. 11:1–11:4.

[10] A. S. S. Sánchez and M. T. Valderrama, "Sonification of EEG signals based on musical structures," in *Proc. Pan Amer. Health Care Exchanges*, 2013, pp. 1–1.

[11] E. Jovanov *et al.*, "Tactical audio and acoustic rendering in biomedical applications," *IEEE Trans. Inf. Technol. Biomedicine*, vol. 3, no. 2, pp. 109–118, Jun. 1999.

[12] A. Ahmad, S. G. Adie, M. Wang, and S. A. Boppart, "Media 1: Sonification of optical coherence tomography data and images," *Opt. Express*, vol. 18, no.10, pp. 9934–9944, May 2010.

[13] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, and S. Yaacob, "A stereo image processing system for visually impaired," *Int. J. Inf., Control Comput. Sci.*, vol. 2, no. 9, pp. 1–10, 2008.

[14] B. Chidester and M. Do, "Assisting the visually impaired using depth inference on mobile devices via stereo matching," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jul. 2013, pp. 1–6.

[15] P. B. L. Meijer, "An experimental system for auditory image representations," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 2, pp. 112–121, Feb. 1992.

[16] M. Capp and P. Picton, "The optophone: An electronic blind aid," *Eng. Sci. Educ. J.*, vol. 9, no. 3, pp. 137–143, 2000.

[17] A. C. G. Martins, R. M. Rangayyan, and R. A. Ruschioni, "Audification and sonification of texture in images," *J. Electron. Imag.*, vol. 10, no. 3, pp. 690–705, 2001.

[18] W. S. Yeo and J. Berger, "Application of raster scanning method to image sonification, sound visualization, sound analysis and synthesis," in *Proc. Int. Conf. Digit. Audio Effects (DAFx-06)*, Sep. 2006, pp. 309–314.

[19] S. Matta, D. K. Kumar, X. Yu, and M. Burry, "An approach for image sonification," in *Proc. 1st Int. Symp. Control, Commun. Signal Process.*, 2004, pp. 431–434.

[20] K. Ivan and O. Radek, "Hybrid approach to sonification of color images," in *Proc. Int. Conf. Convergence Hybrid Inf. Technol.*, 2008, pp. 722–727.

[21] M. Banf and V. Blanz, "A modular computer vision sonification model for the visually impaired," in *Proc. 18th Meeting Int. Conf. Auditory Display*, 2012, pp. 121–128.

[22] M. Banf, R. Mikalay, B. Watzke, and V. Blanz, "Picturesensation—A mobile application to help the blind explore the visual world through touch and sound," *J. Rehabil. Assistive Technologies Eng.*, vol. 3, pp. 1–10, 2016.

[23] M. Banf and V. Blanz, "Sonification of images for the visually impaired using a multi-level approach," in *Proc. 4th Augmented Human Int. Conf.*, 2013, pp. 162–169.

[24] S. Scavaco, J. T. Henriques, M. Mengucci, N. Correia, and F. Medeiros, "Color sonification for the visually impaired," in *Proc. Int. Conf. Health Soc. Care Inf. Syst. Technologies*, 2013, no. 9, pp. 1048–1057.

[25] G. Parseihian, C. Gondre, M. Aramaki, S. Ystad, and R. Kronland-Martinet, "Comparison and evaluation of sonification strategies for guidance tasks," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 674–686, Apr. 2016.

[26] A. Radecki, M. Bujacz, P. Skulimowski, and P. Strumillo, "Interactive sonification of images on mobile devices for the visually impaired," in *Proc. Signal Process.: Algorithms, Architectures, Arrangements, Appl.*, Sep. 2017, pp. 239–242.

[27] M. Mignotte, "A label field fusion model with a variation of information estimator for image segmentation," *Inf. Fusion*, vol. 20, pp. 7–20, 2014.

[28] A. Munsell, "A pigment color system and notation," *J. Psychol.*, vol. 23, no. 2, pp. 236–244, Apr. 1912.

[29] C. Loughlin, *Sensors for Industrial Inspection*. Berlin, Germany; Springer, Dec. 2012.

[30] S. Fausti, D. Erickson, R. Frey, B. Rappaport, and M. Schechter, "The effects of noise upon human hearing sensitivity from 8000 to 20 000 hz," *J. Acoustical Soc. Amer.*, vol. 69, no. 5, Jan. 1981, Art. no. 1343.

[31] T. Elbert, A. Sterr, B. Rockstroh, C. Pantev, M. MÃller, and E. Taub, "Expansion of the tonotopic area in the auditory cortex of the blind," *J. Neuroscience*, vol. 22, no. 22, Nov. 2002, Art. no. 9941.

[32] F. Gougoux *et al.*, "Pitch discrimination in the early blind," *Nature*, vol. 430, no. 6997, pp. 1476–4687, Jul. 2004.

[33] C. Wan, A. Wood, D. Reutens, and S. Wilson, "Early but not late-blindness leads to enhanced auditory perception," *Neuropsychologia*, vol. 48, no. 1, pp. 344–348, Jan. 2010.

[34] Wikipedia, "Piano key frequencies," [Online]. Available: https://en.wikipedia.org/wiki/Piano_key_frequencies. Accessed on: Jun. 13, 2018.

[35] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Comput. Vision*, Jul. 2001, vol. 2, pp. 416–423.


**Ohini Kafui Toffa** received the M.Sc. degree in computer science in 2009 from the University of Sherbrooke, Department of Computer Science, Sherbrooke, QC, Canada. He is currently a Ph.D. student in vision laboratory of the university of Montreal (DIRO) and a Supervisor of Software development at Intrado Montreal. His research includes multimodal methods, applied mathematics and machine learning, in sound, image and video classification.

**Max Mignotte** received the DEA (Postgraduate degree) in digital signal, image and speech processing from the INPG University, France (Grenoble), in 1993 and the Ph.D. degree in electronics and computer engineering from the University of Bretagne Occidentale (UBO) and the digital signal laboratory (GTS) of the French Naval academy, France, in 1998. He was an INRIA post-doctoral fellow at University of Montreal (DIRO), Canada (Quebec), from 1998 to 1999. He is currently with DIRO at the Computer Vision & Geometric Modeling Lab as Professor at the University of Montreal. He is also a member of LIO (Laboratoire de recherche en imagerie et orthopedie, Centre de recherche du CHUM, Hopital Notre-Dame) and researcher at CHUM. His current research interests include statistical methods, Bayesian inference and energy-based models (especially encoding non-local pairwise pixel interactions) for solving diverse large-scale high-dimensional ill-posed inverse problems in imaging.