

Visual textures seen as sonograms. (Is it nice looking and sounding?)

Daniel Arfib

LMA-CNRS
Marseille, France
arfib@lma.cnrs-mrs.fr

Lucas Tamarit

Swiss Center for Affective Sciences
CISA - University of Geneva, Swiss
lucas.tamarit@cisa.unige.ch

ABSTRACT

Using images as sonograms is a dream that is sometimes claimed as being already done. But strange laws exist in sonograms, which are difficult to achieve for images.

We show in this article what kind of laws exist, and how we can bypass them, or at least take in account the difficulties to consider arbitrary images as sonograms. Three solutions are described and tested. Finally the integration of “sonogramical sonification” is inserted in the interaction and gesture domains with an attempt to link visual research on textures with interactive sonification..

1. INTRODUCTION

Transforming a sound into an image is something that the scientific world is accustomed to, and acoustic journals since decades include images from sounds, mostly of them using the Kay sonagraph. Digital implementations have followed, using short Fourier transform, and displaying the magnitude of it.

Transforming an image into sound may reveal more tricky, depending upon the goal of the research: if the goal is to find a sound which sonogram is near to the target image – this is one goal of this article – we must progress by steps.

First if the image is itself a sonogram, so the magnitude of a short Fourier transform, we should be able to reconstruct the sound, because analysis-synthesis programs such as the phase vocoder can do so. But this means that we have to guess a value for instantaneous phases, and keep the same window for resynthesis that was taken for the analysis. Guessing phases is not a trivial task, and section 3 will describe some solutions.

Secondly if the image is arbitrary, there is no chance that we can find a sound which sonogram will be that image. This is due to constraints that sonograms have: the points are not independent, but linked by what signal processing people call the “reproducing kernel” which is a relation between adjacent points. So at that point the idea is to find a good approximation that resembles the original but belongs to a possible sonogram, with associated phases values.

The consequence is that converting images to sounds as if these images were nearly sonograms is not unique, and needs some assumptions that need to be put affront. Some solutions are given in this paper on how render such images in a proper way, namely with three different techniques which all give an answer to the image to sound translation. Nevertheless this is not the end of the story: viewing images as sonograms implies that we have a time and frequency axis, and we differentiate vertical and horizontal direction. As we will see, rotation of images give strong artifacts, and in a general ways there is no insurance that pertinent features of the image will be heard, and reciprocally details in the

image will give rise to prominent parts of the sound. The perception of images is somewhat different from their interpretation as sonograms.

Interaction with this sonification process can be thought in two ways: we can interact with an image, listening to the sonification of the part we are focusing on. We can also go the other way, which means create images in an interactive way – and visual textures are excellent candidates for that – and obtain sounds we can interact with. This last field corresponds to a gesture-controlled audio system, and is very near to digital music instruments.

2. WHY NOT EVERY IMAGE CAN BE A SONOGRAM ?

Sonograms have signal processing constraints: they are normally complex (in the mathematical sense of real and imaginary part also seen as magnitude and phase) and there is a relationship between magnitude and phase as well as a relationship between a point on the sonogram and the surrounding points. So a visual image coming from a synthesis process or from a natural capture has no chance to be a valid image (in the sense it is the transform of a sound): it is usually composed of real values (in the mathematical sense, no imaginary part in this signal) and the relationship between independent points is arbitrary..

2.1. Sonograms

Sonograms were initially obtained through an apparatus named Sonagraph (Fig. 1), in which sound was recorded on a track around a cylinder, and where this sound was analysed with a running filter and a « carbonized paper ».

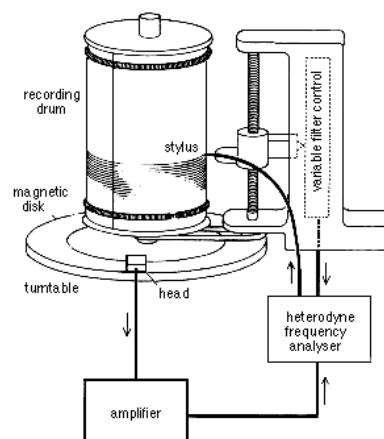


Figure 1. image of a Sonagraph extracted from <http://jproc.ca/rpp/sonagraph.html>

From its start the Sonograph was able to give different interpretations, depending on the bandwidth of the filters: with a short band filter, the sinusoids appear at best, but every transient is blurred, while in the contrary when the bandwidth is widen, the partials disappear and become more a spectral envelope, while the transients appear sharp.

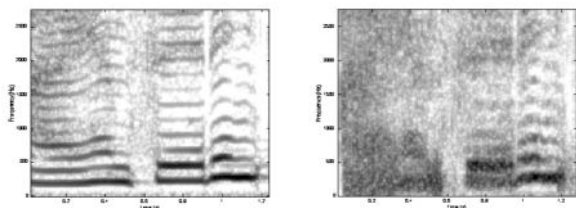


Figure 2. digital sonograms of the same sound with a large (left) and short (right) window..

A digital sonogram (note: a “sonogram” is normally a drawing coming from a Sonograph, while sonograms is the generic term for a time-frequency representation of a sound) can nowadays be obtained via a sliding short-time Fast Fourier transform. The result of this transform consists in values that are put on grid, which consist of the magnitude and phase of the successive Fast Fourier transforms. As the Fourier transform is invertible, the condition for a perfect reconstruction is that the sum of the square of the windows (if the FFT and IFFT are windowed) is a constant. As an example using FFT with a 1024 points Hanning window this condition is fulfilled with a hop size (interval between to windows) is 256 points or its submultiples (128,64,...). There is no unique transcription of a sound to a sonogram: this is extremely dependant upon the shape and size of the window. On figure 2 we can see two digital sonograms obtained from the same vocal sound, with two different window size window size Moreover there is another aspect that comes in when one deals with images: not every image can be the representation of a sound. Adjacent points are linked by what is called “the reproducing kernel”. The easiest way to induce the signification of a reproducing kernel [1] is to take a single point of an image, make a forced inverse FFT reconstruction of a sound and take again the sonogram of it. We see then not only a point but a spot, which is the reproducing kernel. We can say that the image is blurred, and this is what we see on time-frequency images: they are either sharp in frequency or in time, but not both. True sonograms are unaltered by such a manipulation (image -> sound -> image) while arbitrary images are.

2.2. Phasograms

Sonograms are usually considered as the magnitude of this transform, and it is interesting to see how a phasogram can be paralleled to it. The term “phasogram” has been introduced in the eighties and is the time-frequency representation of the phase values of the sliding FFTs [2]. The phasograms are different according to the window size (Fig. 3), and there is a kind of link between magnitude and phase, they cannot be arbitrary. As the phase is rapidly changing, this phasogram representation is best viewed with a short hop size otherwise it appears as sampled and moiré patterns can appear (which does not bias the perfect reconstruction). So, in short, to reconstruct a sound form a given magnitude, we have to provide a phase value for each point of the plane. This is not an easy game, and some solutions will be presented in the next paragraph.

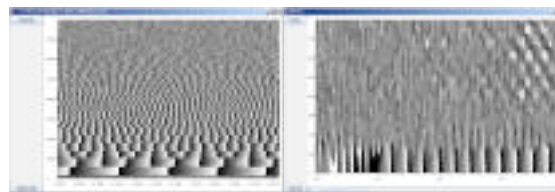


Figure 3. Phasograms of the same flute sound with a large (left) and short (right) window.

3. SOME SOLUTIONS

We will now give three possible solutions in order to “reinvent” sounds issued from images. The first one is to force the magnitude to be close to target values, and try to find successive approximations of phases. The second one consists in separating three sub images via a 2D filtering, in order to reconstruct independently sinusoids, transients and noise. The last one is an exploration of mathematical morphology, which allows a clean distinction of shapes in an image. These three solutions are only expedients but they are nice looking and sounding.

3.1. Iterative processing

Now we are sure of some facts:

- Knowing only the magnitude of a sliding short Fourier transform is not enough to reconstruct the sound of which it is the representation
- The shape and size of the analysis and synthesis window has to be given otherwise it makes no sense talking about a sonogram
- Images coming from the analysis of a sound are valid, while arbitrary images are not: these last ones cannot be the magnitude of a sliding short-time Fourier transform, so their transcription depends upon an interpretation of the information contained in the image

A way to recover a proper phase is thus very important in the case of sonograms, and the idea of an iteration is a good one (Fig. 4): starting from initial values for the phase plane, one can recover a sound (by summing the inverse FFTs) and we can get a new sonogram. If we impose again the target magnitude and keep the new phase, we initiate an iterative process that may or may not converge depending if the target is “close-to-a-sonogram” or not. The algorithm can be described as such: given a magnitude surface, and arbitrary starting values for phase, perform this loop:

```

n. repeat
    reconstruct a sound from magnitude and phase
    calculate the sonogram (new magnitude and phase)
    keep the initial magnitude and take the new phase
end
    
```

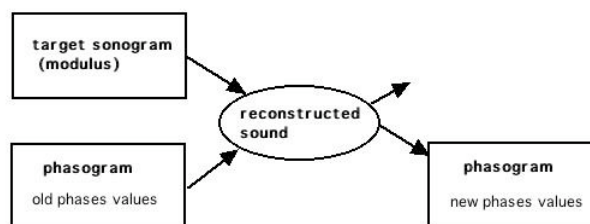


Figure 4. The iteration process: building new phases.

In the case of sonograms coming from original sounds, we recover little by little this sound, but as the ear is very sensitive to defaults, the process is quite slow and experience has proven that 100 iterations provide a quite good result. If the initial phases are at random (between 0 and 2π), the algorithm converges from a breathy (or noisy) sound to the original one. If the phases are put to a zero value, the algorithm converges from a robotic voice (at the frequency corresponding to the hop size) toward the original sound. Other algorithms help providing good starting values. This is the case of an algorithm suggested by Laroche [3] for phase vocoder applications where the phase is developed along spectral lines according its natural frequencies, and this phase is copied around its neighbourhood. This algorithm consists in taking, at a given time, only the bins that contain the maxima of the magnitude, calculate the frequency corresponding to these bins, and so calculate the phase estimation for these bins. Next the interval between two such bins is filled by the values of the adjacent bins. The idea is to provide a “natural variation of phase”, and preserves the fact that a point is linked to its neighbours. We will see in next paragraph the implementation of such an algorithm, but for right now it is only useful to know that algorithms exist that suggest good values as a starting point.

When it comes to an arbitrary image, the algorithm does not converge towards the exact sonogram, but to a solution that is valid and not too far from the original image. But no claim can be then given that we have “sonified” the image via a sliding short Fast Fourier transform.

3.2. A three filters solution

So it is now clear that we must make an interpretation of the image in order to get a sound. The computer music literature offers us a point of view which may reveal interesting: modelling a sound is often to take in account the perception, and particularly the sinus-transients-noise models offers a good point of view: sinusoids are represented in sonograms by horizontal lines, whenever their frequency does not change too fast. Transients are represented by vertical lines and noise by random aggregates of random values. Even more interesting is the fact that the phasogram associated with sine, transients and noise is clear: sinusoids are accompanied by a band of phase turning at the same frequency (in fact it is the phase of the different partials), while transients give phases that correspond to the analyzing frequency (that of the bin) and give a tree shape around the transients. The rest can be considered as noisy, in which case a random value for the phase can be a good starting point.

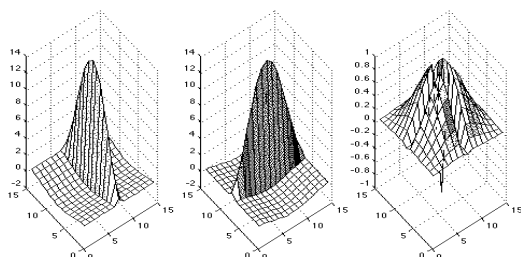


Figure 5. Representation of the three 2D FIR Filters.

The idea is then to subdivide the image into three sub images, which correspond to these three components, and to try to allocate phases corresponding to each of these three

distinctions. This is a very strong interpretation of an image seen as a sonogram: it means we will privilege the horizontal and vertical lines of the image, giving them the meaning of sine waves and transients, and keep the rest as residual noise.

The simplest idea is to use three complementary filters (Fig. 5), and the obvious thing is to use patterns that we are looking for in the image (filtering is a 2D convolution with FIR filters). Though not optimal in terms of signal processing, they achieved their goal which is to give three complementary sub images. (Fig. 6 and 7)

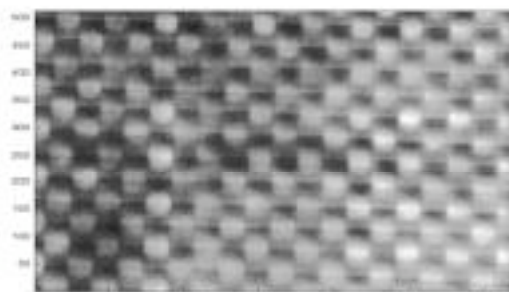


Figure 6. an image to be sonified (Brodatz textures).

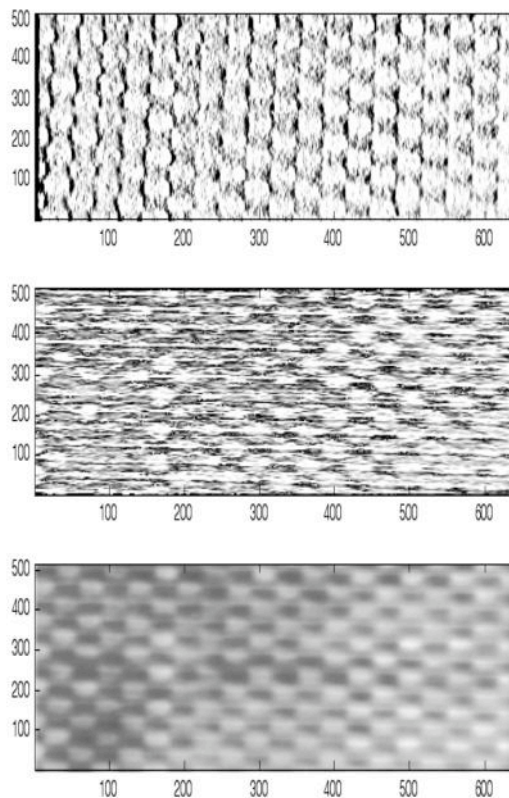


Figure 7. the three subimages resulting from the filtering process.

To reconstruct sounds from these sub images, one processes differently for each of them:

- for the horizontal lines, every successive column (vertical) of the image is processed this way: the maxima are found, an evaluation of the new phase is done corresponding to this maxima (Fig.8).

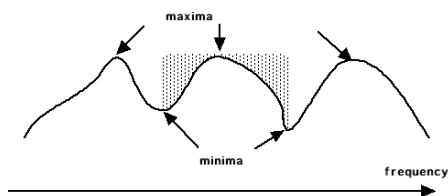


Figure 8. Finding the maxima of a column and the zone of appliance of the new phase.

An algorithm similar to the one used by SMS (Sinusoidal Model System, X. Serra [4]) evaluates a parabola around this maximum, so the frequency fr of the sinusoid, and then unwraps the phase by adding to the precedent value of the phase the interval of phase $\Delta\phi$ corresponding to the frequency of this sinusoid.

$$\Delta\phi = 2\pi * \Delta * \text{hop} * fr / SR \quad (1)$$

$$\phi(n, \text{col}) = \phi(n-1, \text{col}) + \Delta\phi \quad (2)$$

where fr is the estimated frequency, hop the hop size (interval in samples between two columns, and SR is the sampling rate.

Then this phase is duplicated in all the rows up and down till either the half sum of two maxima (Laroche's way), or till the minimum in between these two maxima (Fig.8). Thus the entire plane can be filled by phase, and a reconstruction of a sound can then be done, using or not the iteration algorithm provided before.

- for the vertical lines (second filtering) one processes by rows, and each maximum is pinpointed as having a zero phase. Next for one row all the other columns are filled by values corresponding to a "natural frequency" (that of the bin) which means that a interval of phase is calculated

$$\Delta\phi = 2\pi * \Delta * \text{hop} * f / SR \quad (3)$$

$$\phi(\text{row}, n) = \phi(\text{row}, n-1) + \Delta\phi \quad (4)$$

A sound can then be reconstructed, with or without using the iteration algorithm described before.

- For "the rest", as it is interpreted as noisy signal, it is better use random values to fill the time-frequency plane and let the algorithm find his way in reconstructing phases.

The sonic process consists then in a mixing of these three sounds, which will render what is seen on the image interpreted as a sonogram. If the image is the sonogram of a sound, the process is quite clear, and the result approaches the original sound. If the image is not a sonogram, one obtains sounds that are better resembling to what a human being would think in a "sonogram reading" process.

We must be very clear at this point that vision and audition can be paralleled, but not made equal: vision of image is 2D, and though horizontals and verticals have a meaning, vision is looking for shapes that are 2D. Audition is very strong in the fact that it needs time, so that the vertical and horizontal axis are given two very different meanings. Moreover the emphasis on sinusoids means that we are looking for horizontals, and each slide of a small angle of an image will give a glissando. So let us stay modest, and remember that the focus of this paper is on "is it possible to interpret images as sonograms", which is a very strong limitation.

3.3. Other ways: the morphological approach

Among many other methods that can detect lines in an image (such as the Hough transform or the Radon transform) we can choose an approach based on mathematical morphology.

Principles and algorithms of mathematical morphology allow the choice of criteria directly related to the shape and dimensions of objects to be detected. In our case our interest is to emphasise horizontal and vertical lines contained in images, and we naturally turn to the operation named morphological opening, a complete description of which can be found in [5].

Morphological opening (fig. 9) needs the definition of two parameters: an image (Black and white, or grey levels) and a structuring element. The process finds the interesting objects, with a two step process:

- the original image is eroded by taking, for each pixel, the minimum value of the image covered by the structuring element centered on this pixel.
- Then a second process dilates this eroded image this time by taking the maximum value of the image covered by the structuring element centered on this pixel. This operation is dual to the erosion.
- In short this keeps the regions where the structuring elements can be included and discards the other regions; this preserves the relevant features of the image and skips the non pertinent ones.

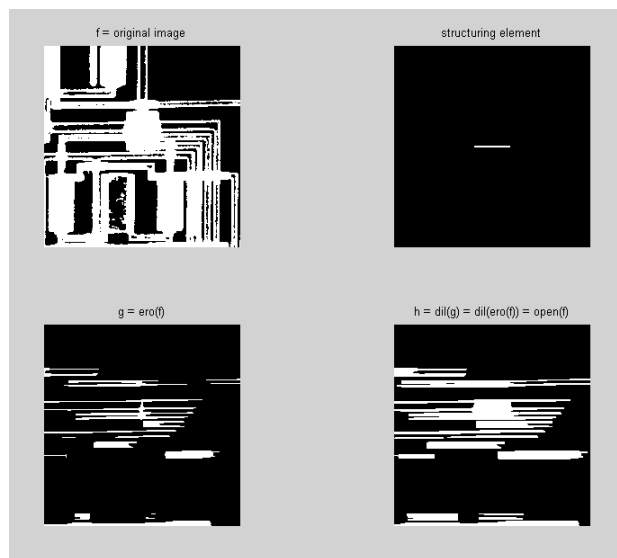


Figure 9. mathematical morphology: given an image and a structuring element (upper) the image is eroded (lower, left) and then dilated (lower, right) .

As an example with an image with squares and disks, using a disk as a structuring element will discard the squares and let only the disks. The big difference with 2D FIR filtering seen in 3.2 is that it is a non linear operation, that clarifies the image using specific shapes.

In our case, the structuring elements are horizontal and vertical segments, and as it is exemplified (see Fig. 9) this gives really interesting results in terms of vision process.

Once again there is an interpretation of what should be seen as a sonogram in an image, and the morphological mathematics are precise, concise, and can be of good help to verify the hypothesis on which one bases him/herself to

sonify an image. When these new images are created, it is up to the sonification process to bring them to sound, and as this mathematical morphology give also subimages, it is good practise to use the same algorithms as in 3.2.

3.4. Mathematics and perception

Using images as if they were sonograms brings many artefacts. The perception of images and sounds differ and the sonogram interpretation is strongly different from a regular “visual” interpretation:

- horizontal lines in a sonogram give rise to harmonics (sinusoids) and vertical to transients. This means that these two directions are privileged in the sonification and correspond to what we see in a weaving process

- the sonic process is extremely dependant upon rotations: a small rotation of an image made of horizontal lines will make the sound pass from affixed pitch sound to a glissando (sliding pitch) while the image seems alike.

- The three solutions given in section 3 give different interpretations: the first one tries to fit at best the target magnitude, in a mathematical sense; the second one (three filters) separates the sonification in three, and applies a different sonification to the three parts, very near to what the ear is doing. The last one, which uses a shape detection will enhance the strong points of the image and erase small details.

4. PERSPECTIVES AND WIPS

We now consider some interactive process (gesture-controlled synthesis and sonification) in the building/interpretation of such textural images. They have to be considered as works in progress, but it is always at the beginning that sparkling ideas appear.

4.1. Algorithms from the visual domain

The initial goal of this research was to see how the methods to create images and sounds can be compared and if some useful links can be made between the two (see [6,7] for a list of references). The subject of textures is the one that is focused, with a question: we know how to create visual textures, can we use the same techniques to make sonic textures. Image synthesis and processing offers techniques, two of them being good candidates to be linked with the sonogram approach developed before

- procedural techniques allow the making of textures from scratch which can then be sonified.. The idea is not new: some programs such as Metasynth [8] or Audiosynth [9] use this metaphor in order to make sounds, using a forced reconstruction. The novelty here is to try to get images that are devoted to sounds, in a word making procedures to draw sonograms or pseudo sonograms that can be interpreted as such.

- germination techniques allow the making of a visual texture starting from a sample of it. The idea (WIP, work in progress) is to make a sonogram of a recorded sound, better be a sonic texture, and then apply germination techniques to create a running sonogram [10].

Interaction may come in two points:

- Interaction in the interpretation of the image means that we change parameters in this interpretation of an image that is predefined (for example scrubbing an image, but also changing the parameters of the interpretation: window length, image orientation, zoom, aso). This way we really are in a situation where the interaction is with the image process

and where classical manipulations (pointing, zoom, rotation) can be linked to gestural devices. Good candidates for that are tangible devices, which allow a direct interaction with image processing.

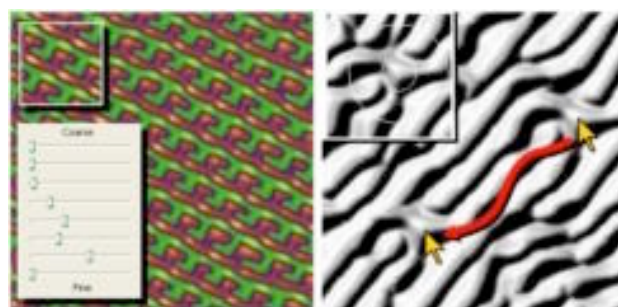
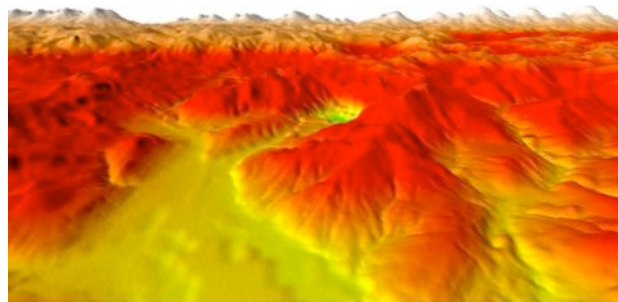


Figure 10. Landscape generation and real time interaction with textures (Lefebvre and Hoppe)

- Interaction with an image creation means interaction with its sonic counterpart, providing the fact that we tolerate a “scanning approach”, as the sonic textures are developed across time. Such interactions are oriented towards the musical side: one can trigger the sonic scanning of images rerecorded in a bank, this making a perfect organ instrument, with facilities of control linked to the diverse interpretations of sonograms (for example time stretching is made easy). Conversely, we can compute images “on the fly”, and render them in real time. This track that we currently follow consists in deviating the visual algorithms in order to create new columns in a sonogram. As an example a short sonogram of running water can be completed via texture making algorithms (sometimes named landscape on the fly generation, see Fig. 10) and sonified in real time. Good examples of such techniques can be found in [10]. Their work also includes the manipulation of textures in real-time which can be an interesting track for interactive sonification.

4.2. gestural control

Adding a gestural device in the interaction loop is very near to the concept of a digital music instrument: this is not new, but relatively few works have been done on the concept of a digital music instrument producing textures. The notion of “malleable virtual objects” is the one developed with JJ Filatriau (Fig.12) using the exploration of visual textures or displacements of fractal trees [11].

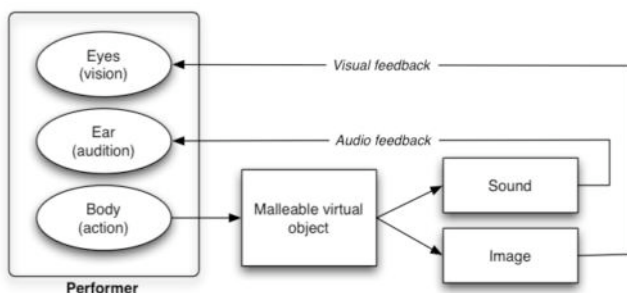


Figure 11. A gesture linked to a virtual malleable object.

These objects are not on the gestural side (and so are not malleable devices [12]) but are on the sonic part, as a top level description of sounds. They are virtual objects (Fig 11) that have a visual and a sonic behavior that are linked, and when done properly this appear as a cross-modal object.



Figure 12. A gesture linked to a malleable virtual object.

Sonograms could also be a good candidate to be a malleable virtual object, providing that we work more at the meso level (the one of events) rather than at the micro level (the one of timbre): Sonograms could be produced according gestures, but their sonification is triggered by other gestures. In fact we have deliberately to separate the two phases: the creation of the image and its interpretation, but they can be sometimes mixed in which case we can run at the microlevel of sound in real time.

5. CONCLUSION

This article shows that links between vision and sound are not as obvious as one could say. As a matter of fact using images as sonograms usually give many artefacts that can also be considered as parts of the creative process in music. But it is also important to try to take them in account in order to do a proper sonification of images as if they really were sonograms. Then interaction and gestural control can play the game.

6. REFERENCES

[1] Kay Sonagraph, <http://jproc.ca/rrp/sonagraph.html>
 [2] D.Arffib, F. Keiler, U. Zoelzer (2002): ""Time-frequency processing", in " DAFx digital audio effects ", pp 237-292, ed U. Zoelzer, Wiley and sons
 [3] J. Laroche and M. Dolson, "New Phase-Vocoder Techniques for Pitch-Shifting, Harmonizing and other exotic Effects." *Proceedings of IEEE Workshop on*

Applications of Signal Processing to Audio and Acoustics, 1999
 [4] Serra&al: Spectral Musical Sound Modeling With Sinusoids Plus Noise [published in C. Roads, S. Pope, A. Picialli, G. De Poli, editors. 1997. "Musical Signal Processing", Swets & Zeitlinger Publishers]. See also <http://www.iaa.upf.es/~sms/>
 [5] P. Soille Morphological Image Analysis : Principles and Applications,,Second Edition, Springer
 [6] G. Strobl G. Eckel, D. Rocchesso, "Sound texture modeling : A survey," in Proc. of the Sound and Music Computing Conference (SMC'06), Marseille, France, 2006.
 [7] Arfib D., Couturier J-M, Filatriau J-J. "Some experiments in the gestural control of synthesized sonic textures", in Gesture in Human-Computer Interaction and Simulation, Lecture notes in Artificial Intelligence, LNAI 3881, eds : S. Gibet, N. Courty and J-F. Kamp, publisher : Springer Verlag, 2006.
 [8] <http://www.uisoftware.com/MetaSynth/>
 [9] Audiosynth <http://www.audiosynth.com>
 [10] S.Lefebvre, H. Hoppe, Parallel controllable Texture synthesis, siggraph 2005
 [11] Filatriau J-J, Arfib D., "Using visual textures for sonic textures production and control", in Proc.eedings of the 9th International Conference on Digital Audio Effects
 [12] Matthias Milczynski, Thomas Hermann, Till Bovermann, Helge Ritter : A Malleable Device with Applications to Sonification-based Data Exploration , London, UK, Proceedings of the International Conference on Auditory Display (ICAD 2006), 69--76, Eds.: Stockman, Tony, Department of Computer Science, Queen Mary, University of London, 6 / 2006