PERCEPTUALLY MOTIVATED SONIFICATION OF MOVING IMAGES

Jean-Marc Pelletier

Keio University
Graduate School of Media and Governance
5322 Endo, Fujisawa, Kanagawa, Japan
jovan@sfc.keio.ac.jp

ABSTRACT

In this paper a framework for translating moving images into sounds is described. Starting from hand-drawn optical soundtracks in the 1930s to recent computer vision-based approaches several attempts have already been made in this direction. However, in many cases, the relationship between the visual and sonic forms is not quite clear. For this reason, the framework presented here uses Gestalt principles as guidelines for the mapping of local image descriptors to sonic components. A representation of the movement in an image sequence, sampled at characteristic feature points is used to control an equivalent set of sonic components, either continuous tones or discrete grains. It is shown that by fostering perceptual integration of sonic components in a parallel fashion to how the image features are integrated, inter-sensory similarity can be achieved.

1. INTRODUCTION

1.1 Obective

Nature has always acted as a great force of inspiration for countless artists and musicians. Sensing technology allows us to seek music in the silence of physical shapes and phenomena. In particular, imaging sensors provide us with rich information, but meaningful patterns found in images must be extracted through analysis before they can be used musically. The objective is thus to propose a framework for transforming visual patterns into sound – in other words, sonifying – in such a way as to preserve a certain degree of similarity between the resulting sound and the original images.

1.2 Sonification

The formal definition of sonification is “the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation.” [6] In an artistic context, however, the notions of communication and interpretation are rather vague and may be replaced by “artistic expression.”

1.3 Image Sonification

The sonification of images, thus, involves translating relationships between image data (pixels) into sound. While this is not explicitly stated in the definition of sonification, there is an assumption that this is performed in an iconic fashion – rooted in forms rather than semantic meaning.

1.4 Perceptually Motivated Sonification

The data relationships that are translated into sound need not form immediately evident forms and structures in their original state. As a matter of fact, in many cases sonification is used to help identify patterns in data that would not be otherwise noticeable [5]. However, in an artistic context, it may be desirable to maintain a certain level of similarity between the image and its sonification. In other words, an artist may want to create music that sounds like it looks. In order for this to be achieved, it is important to take into consideration the way both sounds and images are perceived. This is what is meant by “perceptually motivated.”

2. BACKGROUND

The conceptual roots of image sonification can be traced back to the poet Rainer Maria von Rilke, who in his 1909 text *Primal Sound* [10] suggested using a phonograph needle to seek sounds in the lines of the material world to transform experience "in another field of sense." Optical film soundtracks offered the first practical means of mechanically deriving sound from images and as early as 1929, the soviet animator Mikhail Tsekhanovsky wondered if lost music could not be heard by photographing ancient Greek and Egyptian ornaments onto a soundtrack [12]. Soon later, the film-maker Oscar Fischinger explored relashionships between shape and sound in his *Tönende Ornamente* (1932) [8]. From the 1970s onward, video and computer processing have greatly facilitated experiments in image sonification. The composer Yasunao Tone is especially notable in this field, having realized several works such as *Voice and Phenomena* (1976), *Molecular Music* (1982), and *Musica Iconologos* (1993) [1].

3. SOUND, IMAGE AND MOVEMENT

While images can be both static or dynamic, sound, and music in particular, is inherently dynamic. Hence, if we are to seek to create perceptual links between hearing and sight, the sonification of moving images, or image sequences should be more suitable. This not only solves the problem of mapping time to a particular visual dimension, it makes it also possible to work with real time input, by using cameras instead of pre-recorded images. Furthermore, and perhaps more importantly, psychological research shows that motion plays a crucial role in multisensory integration – that is, motion is a central clue for associating a particular sound event to a given visual object [11].

4. IMAGE FEATURES

4.1 Corners and Vectors

Most real world images exhibit a great deal of spatial redundancy, which is what makes image compression algorithms possible. It should thus be possible to reduce images to a set of feature points that exhibit low spatial redundancy. The identification of such points, sometimes called *corners*, is often a first step in many high-level computer vision algorithms. Other types of features, such as lines, can also be identified but corners are advantageous both because there are several efficient algorithms for their detection [7] and because their highly local nature is the most suitable for the method presented here. Corners also typically correspond to perceptually salient points, which is a fundamental requirement for this framework.

By identifying strong corners, we are thus able to reduce an image containing hundreds of thousands of pixels to a set of a few dozen or at most a few hundred points that nevertheless preserves most of the image structure. Once corners have been identified, it is possible to use a tracking algorithm to identify its displacement from one image to the next, yielding a set of vectors (x, y, α, Θ) representing the position of a corner (x, y), as well as its displacement amplitude, or velocity (α) and angle (Θ). This set is the *motion flow field* [9].

4.2 Features and Gestalts

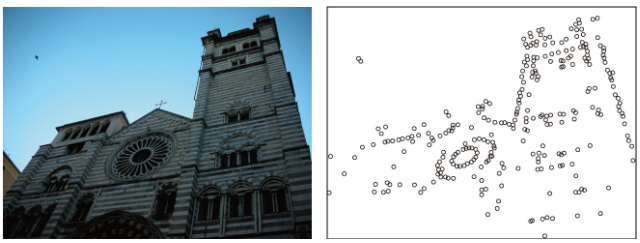

Figure 1: An image and its 283 corners

If we take but a quick glance at the corner map in figure 1, we can identify a number of features such as edges and two concentric circles towards the centre. We may even identify the object represented as a church. However, this image is composed of only a relatively small number of identical circles. All the other objects that we see in the image are thus the result of perceptual grouping. This observation, of course, is one the basis of Gestalt psychology, which identifies a number of laws describing how features are grouped in larger forms, or Gestalts [4].

In a static set of corners, such as fig. a, laws of *proximity* (features close to each other are grouped together), *closure* (gaps between features are closed to form paths) and *continuity* (features are grouped to form the smoothest paths) are at play. If we are to look at the motion flow field instead, the law of *common fate* (features moving in the same direction are grouped together) becomes important.

4.3 Visual and Sonic Gestalts

Gestalt principles can also be used to describe and predict groupings in auditory perception. As a matter of fact, as pointed out by Bregman [2] and others, auditory perception is in this regard analogous to visual perception – visualisation of sonic or musical structures will often be perceptually segmented in parallel fashion to the sound represented.

If Gestalts can somewhat be preserved when representing sounds visually, it stands that the opposite should be true: visual forms can be translated in equivalent sonic structures. Hence, if we are to sonify images in a perceptually meaningful way, care must be taken to preserve some level of relationship between elements when transitioning from the visual to the sonic world.

5. FROM VISUAL FEATURES TO SOUND

5.1 Synthesis Techniques

The motion flow field is composed of highly local image descriptors. Global shapes, as has been shown, are preserved as perceived relationships between these local points. Such shapes can only be perceived if a sufficiently large number of features are present. (The actual number of required features depends on the complexity of the image.)

The most natural way to sonify such a data set is then to use synthesis techniques in which a large number of simple components are added together to form larger and more complex structures. There are many ways in which this can be done, but these methods may be classified as *additive* processes, where a variable number of continuous sounds are modulated, and *granular* processes, where short, static sound events – grains – are generated.

The precise choice of synthesis technique is left to the composer. There will never be a single "correct" way of

sonifying a given image, and so the framework must allow a certain level of freedom to be of creative value.

The framework presented here is thus an approach to control rather than synthesis itself. As such, it provides an interesting solution to the problem of controlling large sets of parameters in some forms of sound synthesis.

5.2 Mapping Parameters

The approach to mapping visual to sonic parameters must be kept fairly simple. If the mapping is too indirect, it is likely that the relationships that exist between the visual elements will not be translated into the sonic realm. Furthermore, each motion vector will always find itself assigned to a single sound component. That way, the number of simultaneous sonic elements will depend on the complexity of the image. Visually dense images will tend to produce denser sounds.

5.2.1 Space

The easiest parameters to map are those that pertain to location. When working in a stereo environment, we can simply pan each sonic component according to the x axis coordinate of the corresponding motion vector. In a surround setting, we can directly map the position of the vector to that of the sound. This will allow complex yet coherent trajectories to be intuitively generated, making this an efficient approach to spatialization. Because the motion flow field informs us not only of the position of visual features but also their movement, it is also possible to simulate Doppler shift effects.

Since we are mapping space to space, most of the relationships that existed between the visual features are maintained. Clustering of sonic components occurs primarily through proximity and common fate, with spatial trajectories more than absolute position playing an important role. It is important to note that due to the nature of multispeaker playback, under certain situations sonic Gestalts may differ markedly from visual ones. For example, two otherwise identical sonic components positioned at either end of the stereo spectrum will be perceived as a single object at the centre while such confusion will never occur with visual features.

5.2.2 Dynamic profiles

While the assignment of position to position is self-evident, other mappings are somewhat more arbitrary. Nevertheless, parameters affecting the amplitude of the sonic components can usually be mapped in a straightforward and significant way.

The most obvious approach is to map the velocity of the motion vector to the amplitude its corresponding sound component. While there is no strict reason for such a relationship to exist, it is not entirely unnatural either. From the bowing of strings to the striking of percussion instruments, fast performance gestures are often associated with louder sounds.

Because the purpose of the framework proposed here is to express in sound visual structures with an emphasis on movement, it follows that motionless features should be silent. Having amplitude envelopes follow the dynamic motion contours establishes a very strong common fate relationship not only between the sound components but also fosters inter-modal grouping between sound and sight.

In some cases, the composer may not wish to use visual motion to control amplitude. This may be because the image contains very strong features that are often static and the composer wishes these too to be expressed. A number of solutions are still available. It is possible, for instance to use the brightness of the image at the feature position. This may prove a good approach in situations where there is a lot of purely temporal motion, for instance blinking lights. However, it is generally advisable to use image elements that change over time in a distinctive manner if perceptual groupings are to carry over to sonic elements as the law of common fate is the most important factor at work here.

5.2.3 Pitch and trajectories

From neumes to piano roll notation, the position of graphical elements has often been used to express pitch relationships. In English words like “high” and “low” are used to describe sound frequency. Spatial trajectories can thus be considered somewhat analogous to pitch modulations or melodies [2]. The relationship between space and pitch, however, is highly arbitrary. There is no good reason to chose the *x* axis over the *y* axis to represent pitch. Neither can we argue that left should be “low” and right “high,” or up “high” and down “low.” Nevertheless, if we wish to preserve relationships between elements during sonification, assigning position to pitch is somewhat motivated.

Clustering through proximity does not function in sound the same way it does in sight. Depending on the harmonic relationship, two frequencies that are close in pitch may not be perceived as belonging to the same object while frequencies that are far apart may be considered to belong together. However, here again, common fate acts as a powerful agent. It has been demonstrated that parallel motion of frequency components acts as a strong unifying force – for example, micromodulation of frequency components causes separate harmonic components to fuse into a single perceptual sound object [3]. Since many real-life objects exhibit a fair amount of rigidity, image features that belong to the same object will exhibit strongly correlated trajectories. When those trajectories are translated to pitch glissandi, the resulting effect of tightly-coupled modulation will often result in a corresponding sound object being perceived.

5.2.4 The other dimension

When projecting two-dimensional coordinates to a single-dimension, as suggested above for pitch mapping, symmetry may sometimes pose a problem. Different feature coordinates can produce the same frequency. This may not always be problematic, as independent pitch trajectories are sufficient to cause sound object segregation.

However, we may be faced with two objects moving in similar fashion on either side of the image. Visually, they may be quite distinct (proximity) but sonically they will fuse to a single entity. If we wish to separate them, we need to differentiate them in some way.

In the situation described above, visual clustering is done through poximity but this law does not function straightforwardly in sound. It is then easier to call upon the Gestalt law of similarity to achieve separation. Similarity does not apply to the motion vectors because they represent only image positions that do not in themselves possess a shape. There is no reason, however, for the sound components to lack a distinctive shape, or rather, a timbre.

Thus, it can be useful to use one of the two dimensions to modulate timbre. Practically, this can be achieved in several different ways. In additive processes, each sonic element can be a rich waveform (sawtooth, square, etc.) which is then passed through a filter whose cut-off frequency is controlled by the position of the motion vector. In granular synthesis, timbre can be modulated by controlling the file offset. In either case, the timbral modulation must be done in a continuous fashion so that two visual features close to each other will yield two sound elements with similar timbres.

5.3 Uses

The framework presented here is intended to afford the composer a fair amount of leeway in how it is used. It may be used to create audio-visual works with visuals and sounds following each other tightly. There are many music vizualisers already available, and video jockeys often use music to control video in real-time. The process can be accomplished the other way around with music generated from visuals, or even with both music and visuals influencing each other.

The framework may be used simply as a compositional tool. The composer may create various sonic gestures from images sequence, eventually only presenting the sound and discarding the visual material.

The framework may, of course, be used in conjunction with other means of control. By having only some aspects of the sound coupled to the visuals a richer relationship can be established between the two.

5.4 Sight Reading

For a composer who uses this framework frequently, it is important to cultivate the eye like a musician would normally cultivate an ear. Once the composer has established a number of favorite mappings and sound synthesis strategies, he or she should learn how to "read" sounds into the shapes and motions of the world. In many cases patterns that are visually interesting may not yield interesting sounds, just as interesting sounds may be produced from less interesting visuals. A great part of the sonifying artist's practice is to learn to hear the sonic potential of visual motions.

6. REFERENCES

[1] Ashley, R., Dekleva, D., Marulanda, F., et al. *Yasunao Tone – Noise Media Language*, Los Angeles, USA: Errant Bodies Press, 2007

[2] Bregman, A.S. *Auditory Scene Analysis*, Cambridge,Mass., USA: MIT Press, 1994.

[3] Chowning, J. M., "Computer Synthesis of the Singing Voice", in *Sound Generation in Winds, Strings, and Computers*, Johan Sundberg, editor (Royal Swedish Academy of Music, Stockolm), 1980, pp. 4-13.

[4] Koffka, K. *Principles of Gestalt Psychology.* New York: Harcourt Brace and World, 1935.

[5] Kramer, G. "Some organizing principles for representing data with sound", in *G. Kramer (Ed.), Auditory Display, sonification, audification and auditory interfaces*, Santa Fé Institute, Santa Fé: Addison-Wesley, 1992, pp. 185-222.

[6] Kramer, G., Walker, B., Bonebright, et al. "The Sonification Report: Status of the Field and Research Agenda", Report prepared for the National Science Foundation by members of the International Community for Auditory Display. ICAD, Santa Fe, NM, 1999.

[7] Mokhtarian, F. and Mohanna, F. "Performance evaluation of corner detectors using consistency and accuracy measures", *Computer Vision and Image Understanding*, vol. 102, no. 1, pp. 81-94, Apr. 2006.

[8] Moritz, W. *Optical Poetry: The Life and Work of Oskar Fischinger*. Indiana, USA: Indiana University Press, 2004.

[9] Pelletier, J.M. "Sonified Motion Flow Fields as a Means of Musical Expression", in *Proceedings of the Internation Conference on New Interfaces for Musical Expression, Genova, Italy*, 2008. pp. 158-163

[10] Rilke, R.M. *Primal Sound and Other Prose Pieces.* Cummington: Cummington Press, 1943.

[11] Soto-Faraco, S. and Kingstone, A. "Multisensory Integration of Dynamic Information", in *The Handbook of Multisensory Processes*, G. A. Calvert, C. Spence, and B. E. Stein, Eds. Cambridge, Mass., USA: MIT Press, pp. 49-68, 2004.

[12] Yankovsky, B. "Акустический синтез музыкальных красок", *Kinovedcheskie Zapiski*, no. 53, pp. 353-367, 2001