Principles for Shape Sonification

THOMAS HERMANN
Ambient Intelligence Group, CITEC – Center of Excellence in Cognitive Interaction Technology
Bielefeld University, Bielefeld, Germany

ABSTRACT: This commentary starts with a critical reflection on Jensenius and Godøy's sonomotiongrams as a sonification technique to represent movement shapes. Based on this we propose alternative mappings that require less information reduction. Furthermore, design criteria such as invariance, convergence, and stability are presented and applied to sonomotiongrams. Finally, we formulate necessary conditions for sonifications of movement shapes to support the perception and categorization of shapes, and we propose an experimental procedure to assess and compare movement shapes from auditory representations.

Submitted 2013 May 15; accepted 2013 May 28.

KEYWORDS: *sonification, parameter-mapping sonification, invariance*

JENSENIUS and Godøy introduce sonomotiongrams as an approach to translate videos to sound so that movement shapes of persons shown in the video result in a systematic way in auditory shapes for shape analysis by listening. The method is clearly introduced as an objective, well-defined, systematic, reproducible algorithm, and thus meets the necessary conditions for a *sonification* (cf. Hermann, 2008), which can thus be used as scientific representation of information.

FROM MOTIONGRAMS TO SONOMOTIONGRAMS

The pre-processing steps such as computing an enhanced monochromatic (intensity) image and the absolute difference image with subsequent thresholding and noise removal to extract 'where the action is' are conceptually straightforward, and they furthermore implicitly implement a principle that is increasingly accepted in interactive sonification: the focus on *excitatory sonifications* which are driven by changes and thus are silent (or: fading into silence) in lack of activity. Particularly for real-time (i.e. online) sonification, this mapping avoids irritation and annoyance for the users as they experience the sound as tightly coupled to (their own) action. What is lost in this translation, however, is the awareness of *static* postures and of patterns that involve invisible sustained muscular tension as part of their definition as these just disappear in the sonification. Yet the authors make clear that their technique will not be able to unlock all possible movement shapes as audible shapes.

A critical point in the definition of sonomotiongrams is the particular choice of how the intrinsic two-dimensionality of video frames is reduced to a one-dimensional vector (of spectral intensities) – which is then basically represented using a pitch mapping. The correspondence between motiongrams and a short-term-Fourier transformation (STFT) is intriguing and in itself suggests the chosen representation of linearly organized spatial/visual activity as spectrally organized sonic activity. But this 2D-to-1D reduction is neither deeply principled nor necessary. For instance, an organization of spatially distributed movement activities into 3D voxel cells, where two axes represent the x (horizontal) and y *(vertical)* image axes and the third represents the time axis, would be straightforward. For example, a rather intuitive mapping would then be to map the x-coordinate to one of N loudspeakers distributed azimuthally from left to right, and to use the y-axis for pitch, resulting in spatially moving pitched sounds. Similar spatio-spectral mappings have been explored in various other sonification contexts such as the sonification of EEG data from spatially distributed electrodes on the scalp (Hermann & Baier, 2007, 2013).

An alternative approach, which sticks to a 1D reduction, is to use Hilbert curves, i.e. space-filling curves that scan the 2D image using a 1D-line. They have the characteristics that nearby points in the image are represented – as well as possible – by nearby points on the curve. Hilbert curves are structurally similar to 1D-Kohonen maps (e.g., see Ritter, Martinetz, & Schulten, 1992) when trained on a uniform 2D distribution. A video sonification using Hilbert curves was introduced by Grond (2007).

Concerning the sound synthesis, the sonomotiongrams apply spectral re-synthesis of the motiongram – practically achieved by a bank of time-variant oscillators. This results in sounds where the frequency f is proportional to the y-coordinate of the image. However, a mapping as $f(y) = f_0 \exp(\beta y)$ may be superior as it respects the nonlinearity in pitch perception and grants equal pitch contrast per unit space along the y-axis.

Even with this modification a rising line in the grid-quantized time-spectrum motiongram would result in differently pitched oscillators that fade in and out according to their activation – which is perceptually quite different from a continuous chirp – just as the superposition of two musical notes is quite different from a single tone pitched in between the two. Approximately, the pitch trend is audible, but I suggest considering an implementation that translates spatial movements truly into *continuous spectral movements*, as this may considerably influence the perception of shape. One proposition to achieve this would be to estimate a number of continuous modes along the spectral vector (axis) whose optimal position can be tracked, mathematically speaking, by gradient descent methods in a cost function that measures quantization error, such as in k-means vector quantization. Then naturally the interpolation between subsequent spectral frames as a continuous pitch curve would be possible. This approach has already been used in our preliminary work on traffic data sonification, where vector quantization was applied to learn prototype vectors that move continuously with the vehicles (Hermann & Milczynski, 2005), providing a basis for *continuous* parameter-mapping sonifications. Linking this approach with sonomotiongrams seems attractive.

SONOMOTIONGRAMS FOR BODY MOVEMENTS

Jensenius and Godøy apply their approach to body movements of musicians. This is a good suggestion as sonomotiongrams offer a highly affordable (i.e. low-cost as only a webcam and computer is needed) and non-parametric method for all the various kinds of movements that musicians perform, without requiring specialized musician-specific sensors. However, the camera-based method exhibits several problems inherent to the sensing principle: the limitation to the rather low video frame rate, the inability to represent movements that the system cannot catch, such as hands occluded by the musician's body or instrument, or if an object moves in front of a very similarly colored background. Furthermore, the textures of the object and background strongly affect the sound: for instance a T-shirt with vertical or horizontal stripes will cause the sonification to sound completely different even for identical movement shapes. Furthermore, movements perpendicular to the view ray (e.g., movements towards the camera) will be largely ignored or misrepresented as they may only lead to subtle increase of an object's size. Thus a single camera may be too limited to include sufficient movement shape information, and I would suggest generalizing the approach towards a set of cameras, or even using today's depth cameras to overcome these issues. Fortunately, sounds can be easily superimposed and thus the approach will scale well with sensors.

Another issue is more difficult to address. Sonomotiongrams depend on the object's position, orientation, and scaling: a linear translation or small rotation of the camera affects the audible shape significantly. It would be nice to have *invariance* from such parameters, allowing a more user-centered rather than image-centered acoustic representation. Steps towards that would first be to localize the body axis and then to use *relative mappings* within this stabilized frame of reference. If one assumes movement shapes to be roughly anchored to the body-egocentric coordinate system, the definition and extraction of features that relate to the body's coordinate system cannot be skipped and replaced.

CONVERGENCE AND STABILITY OF SONIFICATIONS

In situations where experts have to invest a lot of time familiarizing themselves with the auditory representation – such as EEG sonification for diagnosis – *convergence* and *stability* are useful features of sonification (Hermann & Baier, 2006). As it is crucial that experts are not forced to re-learn patterns once the recording equipment is enhanced (e.g., doubling the number of channels, or the temporal resolution), we postulated the principle of *convergence*, i.e. that with increasing resolution – in the case of sonomotiongrams for instance frame rate and image resolution – the sonification should converge asymptotically towards a *limes sonification,* so that structural properties remain *invariant*, but converge and stabilize. The current implementation of sonomotiongrams fulfills the convergence property to some degree, but temporal and spatial quantization can still result in significant perceptual artifacts, and a careful consideration of the mapping for how to reduce these effects, and how patterns stabilize with increasing resolution (i.e. number of oscillators), may help to refine the definition of sonomotiongrams towards a standard method.

UNDERSTANDING SHAPES THROUGH SOUND

Unarguably we are capable of perceiving, recognizing and categorizing shapes in all sensory channels, starting from geometric shapes or dynamic shapes (movements) in visual stimuli, through sonic objects and temporal organized sound in the auditory domain, to shapes such as the evolution of smell and taste while eating or drinking. Crossmodal correspondence is an indicative phenomenon of relationships between these processing systems. Both sound and video have multi-dimensional features along which information is organized, and relationships between these are highly interesting, not only for sonification of auditory graphs or sensory substitution applications for visually impaired persons, but also for developing novel analysis methods for shapes as approached with the sonomotiongrams.

Shape characteristics appear in different granularities, for instance as slope, trend, as turning points in a pitch curve, as repetitive pattern in a sequence, or as a motif or even longer acoustic form. The constitution of shapes is a process that starts from an *analogic representation* and points towards the constitution of concepts or *symbolic representations* (cf. Kramer, 1994) that gain stability (and differentiation) as we encounter examples of the stimuli. For instance, a physician will incrementally gain resolution in discerning chest sounds heard with the stethoscope. To best support such processes of establishing perceptual categories, an *invariance* of the mapping is of paramount importance: the invariance over many encountered sonifications assures that the brain can organize the many similarly structured sonic shapes and thereby establish metrics of similarity and distinction. Applied to the sonification of shapes of movements, this asks strongly for well-normalized and *invariant* mappings from the source domain to the sonic domain. Every single user-adjustable parameter may jeopardize the long-term and time-consuming processes of auditory learning. Just as one would not be able to master typewriting if the layout of the keyboard changed daily, *invariance* is the fundamental basis for our sensory systems to start learning. Although certain perceptual variables lend themselves more to our internal modes of shape formation (for instance shapes in pitch are probably more easily established in listeners than shapes within the release amplitude envelope of sounds), I would think that our auditory system would be pretty well able to adapt decently to most sonification types, even if they are suboptimal.

In my opinion, the sonomotiongram approach as a non-parametric and analogic mapping connects quite well with our capability to extract shapes from signals, but for practical applications there are too many sources of variance that can be problematic in the current definition – e.g., effects of object translation, rotation, even clothing, as mentioned above, furthermore the strong dependency of sound shapes on sonification parameters. Thus, for exploring the potential of sonomotiongrams (and also alternative approaches) to support the understanding, recognition and categorization of shapes in movements, I suggest (a) selecting a single movement type, e.g., a specific jump in dance, or a tai chi movement, (b) precisely defining the setup including exact position of the user, distance to camera, background, and (c) recording a large number of movement executions at varying shapes (e.g., different performer's expertise, expression, different variants). Then it requires time for a number of test listeners to build up their own perceptual metrics according to which they could then rate the dissimilarity of pairs of stimuli, either under a visual, auditory, or audiovisual condition. This would be the way to assess and compare the ability of sonifications to enhance understanding of movement shapes. Sonomotiongrams appear to be a well-suited candidate for such fundamental sonification research.

REFERENCES

Grond, F. (2007). Organized data for organized sound – space filling curves in sonification. In: G.P. Scavone (Ed.), *Proceedings of the 13th International Conference on Auditory Display*. Montreal, Canada: Schulich School of Music, McGill University, ICAD, pp. 476-482.

Hermann, T. (2008). Taxonomy and definitions for sonification and auditory display. In: P. Susini & O. Warusfel (Eds.), *Proceedings of the 14th International Conference on Auditory Display (ICAD 2008)*. Paris.

Hermann, T., Baier, G., Stephani, U., & Ritter, H. (2006). Vocal sonification of pathologic EEG features. In: T. Stockman (Ed.), *Proceedings of the 12th International Conference on Auditory Display*. London: International Community for Auditory Display (ICAD); Department of Computer Science, Queen Mary, University of London, pp. 158-163.

Hermann, T., & Milczynski, M. (2005). Videosonifikation am Beispiel von Verkehrsflußdaten (A28 – 01000100100010101011101010101). Sendung 14 (CD) des "Fremder Sender", Haus am Gern, http://hausamgern.ch/satellit, Switzerland.

Kramer, G. (1994). An introduction to auditory display. In: G. Kramer (Ed.), *Auditory Display*. Reading, MA: Addison-Wesley, pp. 1-79.

Ritter, H., Martinetz, T., & Schulten, K. (1992). *Neural Computation and Self-Organizing Maps. An Introduction*. Reading, MA: Addison-Wesley.