Living sound pictures

by Janus Lynggaard Thorborg

Sonic College, Haderslev, 2015

Abstract

In this document I will explore the research on the process of sonifying continuous visual input, discuss mappings of data and dimensions, and finally present a subjective approach to the problem resulting in a prototype.

As technology slowly allows us to move away from statically programmed pieces of music, a whole new era of audio is upon us: Interactive sound. The field of interest here is modulated music.

Introduction and idea

Before the act of recording music was invented, music was bound to the performing artists, in time and place. For some time, it was still bound to static devices such as radios until the revolution of portable media devices. The music, however, became even more static as a result of being recorded and thus frozen in time. Largely due to computer games, however, the trend is reversing. Since the user is actively and non-deterministically interacting with the game, both the visual and the audio output have to be reactive and to some degree procedurally generated.

As humans inevitably transport themselves around and, in the process of doing so, often listens to music meanwhile - the environment is an obvious source of modulation of the music.

Thus, the idea of this project is to transform the visual input of the environment into aesthetically pleasing music, free in time and space. The prototype is an mobile application, using the built-in camera as modulation source, creating sound on the fly. The actual transformation is done through so called *mappings*, which refer to methods of transforming data in one domain to another domain.

Information in visual signals

The immediate problem seems to be, what does living imagery sound like? Does colour matter? The count and shape of objects? Texture? The fact is, visual signals carry a lot more of information than, for instance, sound. At the most basic level, sound is one-dimensional; amplitude levels as a function of time. Visual signals, however, carry a two dimensional matrix instead - over time. To create a 1:1 map between visual and audio signals would be impossible, because you would have to discard two dimensions of information.

Therefore, it is imperative to create a set of mappings. These mappings, in the context of this project, aim to translate the domains in a semantic, intuitive manner. As the next paragraph also shows us, related works differs comparatively in how they translate the dimensional data.

Mappings and approaches

Light, like sound, is but a simple waveform. As Fourier theory tells us, any waveform (or function, really) is completely described through a sum of correlated complex exponentials. One obvious mapping may therefore simply be to translate light frequency to sound frequency. Disregarding the multiple dimensions and assuming a static and completely evenly coloured picture for now, an interesting empirical study was done by Noriko Nagata, Daisuke Iwai, Sanae H. Wake, and Seiji Inokuchi on *non-verbal mappings between colour and sound*[i].

Their approach was based on *coloured hearing*, a subset of synesthesia; a phenomenon in which some modes of perception involuntarily affects others (eg. hearing colour). Deriving a model based off people having such perceived effects, they tried to generalize the model on normal humans.

Tested mappings include absolute hue to key colour, saturation to sound timbre and light frequency to transposition of sound (linear domain translation). Of these, the only tests yielding somewhat statistical significant results was saturation and timbre, that is, subjects generally associated more colourful images to sounds with richer timbres (more harmonics). This will be utilized in the application.

See Through Sound, by Sofia Cavacoa, J. Tomás Henriques, Michele Menguccia, Nuno Correiaa and Francisco Medeiros[ii], is another model and software tool intended to perceive real-time video aimed at visually impaired. It includes a number of interesting approaches, especially in how it deals with dimensions.

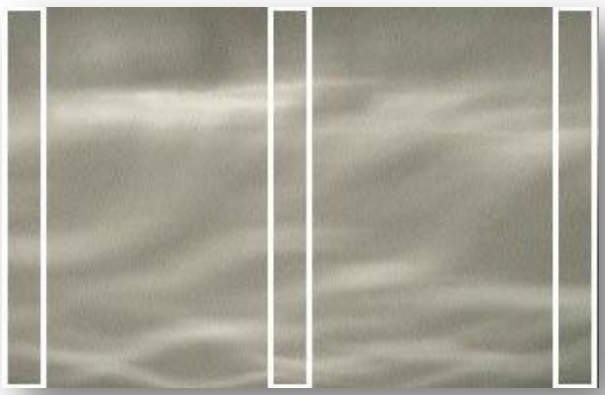
The tool transforms the visual matrix (either real-time video or still images) into rows, which are 'played'. Colour information is decomposed into the HSV (hue, saturation, value) which maps directly onto the sounds. The hue controls the fundamental frequency of the row's sound, while saturation controls timbre and the value controls volume. This is similar to successful mappings from *non-verbal mappings.* There are other domain-specific mappings in the project, namely spatialisation methods. Testing of recognition didn't reveal statistical significant results, however accounting for a confusion matrix of the results yield significantly better results; ie. subjects were able to connect visuals with audio and recognize colours.

One must however consider the approach, test subjects were indeed trained before testing and was learned the specific mappings. The degree of training was found to have a positive result on the tests. The approach can therefore be concluded to not be purely intuitive, but it is of course not critically negative for the target domain.

To conclude the approach, they mapped colours to sound using a HSV model and mapped dimensions to 'time', that is, rows are played in time synthesizing discontinuous music, but still allowing multiple dimensions in real time. This approach will also be used.

A more musical and therefore directly more relevant is a project called *Sound Synthesis from Real-Time Video Images* by Roger B. Dannenberg and Tom Neuendorffer [iii] . They created a synthesizer playing sound generated solely from video input, mapping vertical pixel height to harmonics, where intensity is controlling the volume of the corresponding harmonic. It is therefore an additive synthesizer.

This is a very interesting approach I also will utilize. It still suffers from dimensional loss, as vertical columns are inherently one-dimensional. It compensates for this issue by averaging nearby horizontal pixels and having three 'voices' at once, thereby synthesizing three areas of the image simultaneously (picture from same paper):



As a direct consequence, however, waves propagating horizontally through the image over time will create audible repeated synthesis (perceived as delay-effect), and the synthesizer is actually numb to anything happening in-between the areas.

There are many other relevant works in this area, this was a small summary of work that directly influenced this project - see the appendix for more reading.

Preliminary thoughts

Most people would probably agree that a certain visual environment can be associated with moods (like cosy, scary). Similarly with music. It is important to realize how subjective this perception is, however. Interestingly, for *non-verbal mapping* study, even though a generalized model could not be completed, subjects seemed to associate the same mappings consistently, just differently across the group.

It is therefore also important to keep in mind that while some of my work is technical, the mappings and generated music is subject to *my* perceived relation between sound and colour. It is my intention, however, to generalize the analysing engine to support arbitrary mappings in that the sound generates itself from the available data - sonification - with the idea that people can compose intelligent music pieces for that platform, interpreting the data however they want, thus creating a new musical platform for artists.

A large part of the project was certainly to test the various mappings, and selecting approaches that (in my opinion) worked. I will start by documenting the data the application collects from the image, and afterwards my implementation of the mappings.

Existing prototype

The application works with extending one-dimensional analysis to larger dimensional areas using simple lowpass filters on the orthogonal axis. Realizing not all dimensions can be utilized at once, this is a compromise that is adjustable run-time, which can create interesting effects.

The image is analysed horizontally in the frequency domain utilizing a Fourier transform. The image is also decomposed vertically into RGB components using a similar column as the previously mentioned sound synthesizer. This 'area' transformation is moved horizontally in the image by a sinus translation, possibly creating the same echo effects, but it will cover the whole image eventually. This effect is adjustable in the user interface.

The image is further analysed for global difference in intensity levels for red, green and blue channels (thereby finding deviations from grey values, similar to HSV saturation). This creates a measure of non-grey values in the image. A 'dominant' colour (ternary between red, green and blue) is selected through the means of simple blob-analysis. I will refer to the video describing the product in the appendix for a demonstration of it[iv].

It is also possible to freeze the current rendered image, giving rise to this papers title, with the intention of users being able to both scout around the environment to explore sounds they like in real-time, and possibly eventually fixating on a piece.

Example implementation

The current prototype plays a single, simple musical piece I composed for this specific purpose. The procedural, musical rendering is based off a previous project in audio for a game I made[v], and I will refer to that for implementation. The category of mappings for this song are as follows:

1. The vertical transform controls an additive synthesizer, where red and green values are mapped to the intensity of the height-indexed harmonic in corresponding left and right channels. Blue modulates the frequency in integer harmonic steps, creating fluctuating harmonic notes as the transform translates horizontally. This creates a stereo waveform the synthesizer plays, and this is the basis of the 'chord-instrument' heard in the music. This is also a form of saturation-mapping, as we saw in *non-verbal mapping*, and an implicit intensity mapping.

2. The dominant frequency component in the horizontal Fourier transform is selected and influences the rhythmic real-time composition

such that higher frequency components in the image creates music with more notes, chords and more complex rhythms with smaller divisions, thus mapping (sinusoidal) patterns in the image to rhythm.

3. The global difference in colours control three subtle music effects, where large quantities of green causes a greater mix of a phasing effect, intending to map green to a natural, heavenly state of interaural phasing (very subjective mapping). Likewise, the blue controls a reverb, mapping blue to colder, ethereal and atmospheric sound. Meanwhile, red is in contrast and is completely dry, containing a valve-overdrive effect and a lowpass filter. Red is intended to be a warmer than cold, thus creating effect both through contrast, but also following the notion of valve overdrive to be 'warm', creating even-overtone distortion by asymmetrical soft clipping.

4. The dominant colour controls harmonization of the song, mapping blue to a minor scale, red to a major scale, and lastly green to neither: Green maps to chords missing the third, including also suspended chords. Importantly, they are all in the same scale, so real-time transition is possible.

These mappings combine to create a continuous but constantly changing piece of *music*, derived from the environment.

As a side note, the music piece is kept relatively simple both in sound and composition, as the target platform combined with the prototype system (Unity3D [vi], selected for its platform-independency) is still computationally starved when tasked to do real-time video analysis and audio rendering. The platform is theoretically capable of rendering arbitrary polyphonically complex sounds.

Conclusion and further work

I have now presented a model for extracting usable data from real-time images that can be utilized to generate sound uncommonly utilizing frequency patterns of images, as well as a set of mappings I find intuitive. The approach leans more towards domain transforming instead of translation (eg. giving semantic meaning to colours, instead of just mapping frequencies). More research and testing in this field will definitely be helpful, to hopefully understand how humans map these domains. If further work is done on this project, the next step will be empirical research.

The target platform (mobile devices) carry a lot of other interesting modulation sources, including time of day, weather reports and gyroscopical data. Another interesting source is the microphone input, where techniques such as tempo/beat-following or perhaps noise dampening can be explored. In short, there are many interesting future prospects for interactive audio on-the-go, and a complete set of these might be able to encapsulate and transform complete environmental situations to suitable, intuitive aesthetic music some day.

Appendix

Other interesting projects include:

- *Synaesynth* by Daniel Kerris, a matrix product synthesizing sound from real-time video[vii]
- *Kromophone* by Zachary Capalbo and Dr. Brian Glenney, a product mapping colour hues to seperate instruments (of varying fundamental frequencies) to allow perception of colour through sound[viii]

References

[i] Iwai, D. ; Graduate Sch. of Eng. Sci., Osaka Univ., Toyonaka, Japan ; Nagata, N. ; Wake, S.H. ; Inokuchi, S. "Non-verbal Mapping Between Sound and Color - Mapping Derived from Colored Hearing Synesthetes and Its Applications" in SICE 2002. Proceedings of the 41st SICE Annual Conference (Volume: 1), Aug. 2002, pg. 33 - 38

[ii] Sofia Cavaco, J. Tomás Henriques (Buffalo State College NY), Michele Mengucci (Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa), Nuno Correia, Francisco Medeiros (LabIO). "Color sonification for the visually impaired" at Proceedings of International Conference on Health and Social Care Information Systems and Technologies (HCist), in Procedia Technology by Elsevier (NL) number 9, 2013, pg. 1048 to 1057

[iii] Roger B. Dannenberg and Tom Neuendorffer from School of Computer Science, Carnegie Mellon University. "Sound Synthesis from Real-Time Video Images", 2003. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1508&context=compsci

[iv] Janus L. Thorborg, "Living sound pictures demonstration", 2015, video can be found at www.jthorborg.com

[v] Janus L. Thorborg. "Surfing the Wave" (game) based on the Blunt audio library by same, 2015, implementation and source at www.jthorborg.com

[vi] Unity3D is a multiplatform game development engine. See more at http://unity3d.com/

[vii] The synaesynth can be examined here: http://synaesynth.danielkerris.com/

[viii] The Kromophone project has been discontinued, but can be examined here: http://kromophone.com/